Text-driven Video Acceleration

Washington L. S. Ramos[1]
Dep. de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Email: washington.ramos@dcc.ufmg.br

Leandro Soriano Marcolino
School of Computing & Communications
Lancaster University
Lancaster, UK
Email: l.marcolino@lancaster.ac.uk

Erickson R. Nascimento
Dep. de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Email: erickson@dcc.ufmg.br

***Abstract*—From the dawn of the digital revolution until today, data has grown exponentially, especially in images and videos. Smartphones and wearable devices with high storage and long battery life contribute to continuous recording and massive uploads to social media. This rapid increase in visual data, combined with users' limited time, demands methods to produce shorter videos that convey the same information. Semantic Fast-Forwarding reduces viewing time by adaptively accelerating videos and slowing down for relevant segments. However, current methods require predefined visual concepts or user supervision, which is costly and time-consuming. This work explores using textual data to create text-driven fast-forwarding methods that generate semantically meaningful videos without explicit user input. Our proposed approaches outperform baselines, achieving $\mathbf{F_1}$ Score improvements up to 12.8 percentage points over the best competitors. Comprehensive user and ablation studies, along with quantitative and qualitative evaluations, confirm their superiority. Visual results are available at https://youtu.be/-cOYqumJQOY and https://youtu.be/u6ODTv7-9C4.**

I. Introduction

New digital technologies like smartphones and social multimedia services have made storing and sharing data effortless, leading to a significant rise in data, particularly textual and visual data. Videos have become a crucial medium for people to document their lives and engage socially online. However, lengthy web videos often require fast-forwarding through segments, yet we can still fully understand them. Therefore, techniques must be developed to identify relevant content and reduce the time spent watching long untrimmed videos.

Research on Video Fast-Forwarding and *Hyperlapse* methods [1]–[6] focuses on creating a continuous flow of the video timeline by selecting frames based on camera stability and desired output length. Recent approaches [7]–[18], known as Semantic Fast-Forwarding (SFF) or Semantic *Hyperlapse*, adaptively sample frames based on semantic content, usually splitting the video temporally and using different speed-up rates for each split. The final result is a visually smooth accelerated video with lower playback rates emphasizing the most relevant temporal segments.

The main challenge for Semantic Fast-Forwarding approaches is defining what is relevant to the watcher. Current methods either use predefined *visual concepts* [19] or leverage user-defined semantic labels. Despite their remarkable progress, these methods are limited by a small set of concepts like CARS and TREES or require user-driven supervision.

To tackle these problems, we resort to natural language expressed in texts. In their richest form, texts can be found in pairs with images and videos. For instance, many visual publications on the Internet (*e.g.*, recipe websites and Instagram posts) are often accompanied by text descriptions such as titles, comments, and captions [20]. The overwhelming amount of textual data allows us to leverage visual semantic concepts from text-centric data and create powerful models to solve the Semantic Fast-Forwarding task. Furthermore, it also enables creating methods that fully abdicate direct user supervision.

We argue that natural language expressed in texts carries a latent supervision signal relevant to identifying the visual semantics in video scenes. Hence, we can use text documents from the Internet to determine what is relevant in a video.

Based on these assumptions, in this work, we aim at creating fully automatic text-driven video fast-forwarding techniques capable of using input texts as proxies to identify the most useful video segments. These techniques should correlate the input text with the frames in the original video to infer the most relevant frames to the watcher while dropping the less relevant ones to reduce the video to a target length.

We conducted experiments in two contexts: *i)* Fast-Forwarding First-Person Videos Using Social Network Data and; *ii)* Fast-Forwarding Instructional Videos Using Textual Instructions. In the first approach, we observe that social networks have become an underlying channel for people to interact and express their feelings and opinions. Therefore, personal texts published in such media may contain important cues to infer topics of interest and determine the relevant frames. In the second context, we consider the plethora of online textual tutorials and instructional videos teaching various tasks. Textual instructions are more concise than instructional videos, even though they express the same content. In this regard, textual instructions contain the supervision signal to point out the relevant parts of the video.

Contributions. We can summarize our contributions as: *i*) a novel approach that personalizes a hyperlapse video emphasizing relevant segments according to the user's topic of interest; *ii*) a novel approach based on a reinforcement learning formulation to accelerate instructional videos according to clip similarity scores with textual instructions; *iii*) a model for encoding user and video frames semantics, capable of

[1]This work relates to a PhD thesis.

leveraging raw visual concepts to topics of interest; *iv*) a new Visually-guided Document Attention Network (VDAN) capable of generating a highly discriminative embedding space for textual and visual data; *v*) a reinforcement learning agent that can navigate through videos, adjusting the playback rate based on the semantic load; and *vi*) comprehensive experiments, including ablation and user studies, as well as quantitative and qualitative results.

II. Related Work

A. Video Summarization

In the past several years, video summarization methods were the main approaches for creating visual summaries. Regarding the usage of human annotations, we can broadly divide them into unsupervised, supervised, and weakly supervised methods. Unsupervised methods [21], [22] typically use hand-crafted features or leverage low-level cues (*e.g.*, diversity and representativeness) to identify relevant frames or segments. Supervised methods [23]–[25] rely on human supervision to generate content aligned with human understanding. However, they require human-created summary pairs or fine-grained relevance annotations, which are expensive and time-consuming [26], [27]. Closer to our work, weakly-supervised methods attempt to overcome the difficulty of labeling the frames with relevance scores by collecting information from other sources. They resort to video-level annotations [27], existing summaries (*e.g.*, sports highlights and movie trailers) [26], or auxiliary tasks like moment localization [28]. Despite their success, most summarization methods either ignore temporal aspects or use a relaxed temporal restriction, resulting in visual gaps and breaking the video context.

B. Semantic Fast-Forwarding

In Semantic Fast-Forwarding, the goal is to create a shorter version of the input video that emphasizes relevant content while preserving temporal continuity [7], [8], [29]. Methods for first-person videos, where the camera is constantly moving, must align selected frames to produce a visually pleasing result [9]–[14]; these are commonly referred to as Semantic Hyperlapse. Semantic Fast-Forwarding/Hyperlapse methods roughly fall into two groups: those using predefined semantic concepts as supervision and those using user supervision either before or at runtime.

Approaches using predefined visual concepts typically employ simple object detectors like face [9] and pedestrian [10] detectors, an enhanced set of detectors like the 80 classes from YOLO [13], [14], or low-level hand-crafted features [7]. The most prominent methods in this set are the Sparse Adaptive Sampling (SAS) [13] and SASv2 [14], which model frame sampling as a Minimum Sparse Reconstruction problem using YOLO to construct a content descriptor for the dictionary entries. A major drawback of these methods is their reliance on the accuracy of third-party techniques. Moreover, any application domain changes may need a different detector or intervention from the method's developer.

Some approaches use user-provided supervision to compose frame scores [11], [29]. In recent work, Lan *et al*. [8] introduced the FastForwardNet (FFNet), an RL-based method that summarizes videos on the fly by selecting frames with the most memorable views. Annotated data from video summarization datasets are used as training labels for the agent's reward. In recent extensions, the authors explore distributed and collaborative systems for fast-forwarding multi-view video streams [15], [17]. A drawback of these methods is the need for user supervision to create labeled examples or select objects from a limited set.

In this work, we take a step towards exploring Internet data to infer the importance of frames to the final user in Semantic Fast-Forwarding. Specifically, we leverage the language semantics in freely available raw texts on the Internet to infer the visual semantics in user videos. Using this rich information, we create models capable of defining scenes of significant importance to the watcher.

C. Vision-and-Language Embedding

Cross-modal embedding algorithms have recently emerged as promising approaches for various multimodal tasks. These algorithms create a shared embedding space where features from multiple modalities can be compared. Regarding the density of the information, visual-language methods can represent single sentences along with images [30] or videos [31], and full-text documents (*i.e.*, a set of sentences) along with images [32] or videos [33].

In this work, we also create models that integrate visual and textual domains into a unified space by embedding multiple sentences and frames from the input video. Since these sentences are not generated from the video scenes, we consider them indirect supervision for our methods.

III. Fast-Forwarding First-Person Videos Using Social Network Data

Social networks are key platforms for people to share their emotions, attitudes, and opinions. In this approach, we explore the text-centric data from the users' social networks to create personalized hyperlapse videos. Unlike third-person videos, first-person videos are challenging to accelerate due to natural body movements. Thus, selecting the best frames requires optimizing visual stability, speed, and the semantics.

A. Methodology

Given an input video $V = [v_1, \ldots, v_F]$ of F frames and a document $D = [s_1, \ldots, s_N]$ of N sentences, our goal is to select a subset $\hat{V} \subset V$ with the most relevant frames w.r.t. D, while preserving visual smoothness, temporal consistency, and achieving the target speed-up rate S^*.

Frame Scoring. To score each frame based on user preferences, we build a representation space for visual and textual modalities. We group word embeddings $\mathcal{W} = \{\mathbf{w}_i \in \mathbb{R}^d\}_{i=1}^{M}$ into K clusters using K-Means. Since similar concepts are closer in the embedding space, each cluster/dimension defines a topic of interest, forming the representation vector $\mathbf{x} = [x_1, \cdots, x_K]^\intercal \in \mathbb{R}^K$, referred to as *Bag of Topics* (BoT).

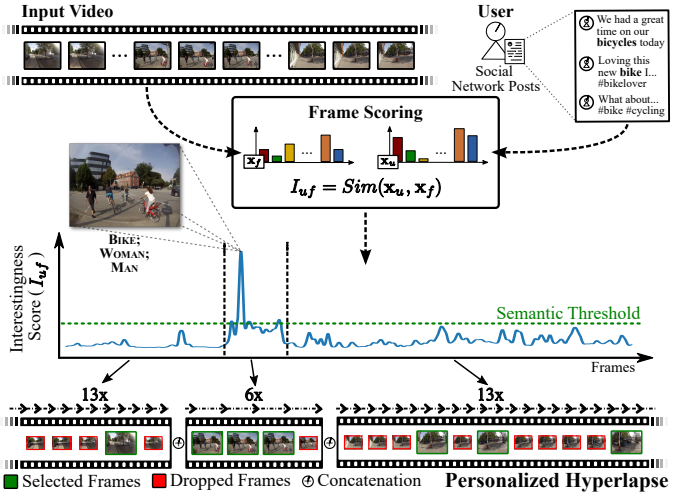

Fig. 1. **Video composition.** First, we calculate a per-frame interestingness score. Then, we segment the video into relevant and non-relevant segments, assigning lower speed-ups to more relevant segments. The final set of frames is a concatenation of the nodes in the shortest path of each segment graph.

We represent a user by first collecting their Twitter posts and extracting the concepts (nouns) from positive sentences. Let $D_u = \{c_j\}_{j=1}^{C}$ be a document composed of C concepts and $\phi : D_u \rightarrow \mathcal{W}$ a function mapping a concept c to a word embedding $\mathbf{w} \in \mathcal{W}$. Thus, given $D_u^k = \{c \in D_u | l(\phi(c)) = k\}$, where l assigns a label $k \in K$ for an embedding, we can represent D_u as the user's BoT $\mathbf{x}_u \in \mathbb{R}^K$, where $x_k = |D_u^k|$.

We represent a frame v_f using weights computed from relevant regions in the scene, extracted using DenseCap [34]. DenseCap produces coordinates, scores, and descriptions $D_f = \{s_1, \cdots, s_R\}$ for the R regions. From these features, we compute the attention (ω_f^a), confidence (ω_f^c), and uniqueness (ω_f^u) weights. ω_f^a weighs the watcher's attention to specific objects or regions, and it is computed as the average pixels value in a saliency map [35]. ω_f^c indicates visual accuracy confidence, and it is directly produced by DenseCap. Some visual concepts may appear only a few times throughout the video, being essential to composing the whole video story. Therefore, we define ω_f^u to assign a uniqueness of a concept, calculated using TF-IDF across the video's document collection $\mathcal{D} = \{D_f\}_{f=1}^{F}$. The final weight for x_k in $\mathbf{x}_f \in \mathbb{R}^K$ is obtained as $x_k = \sum_{r \in \mathcal{R}_f} \omega_f^a(r) \cdot \omega_f^c(r) \cdot \omega_f^u(r)$, where R_f are the regions of interest in v_f.

We use the cosine similarity between $\mathbf{x}_u$ and $\mathbf{x}_f$ to estimate the user's interest in a video frame and produce the Interestingness Score profile for the video (see blue curve in Fig. 1).

Hyperlapse Composition. To compose the hyperlapse video, we employ the algorithm proposed by Silva *et al.* [12]. It splits V temporally into T segments based on a semantic threshold that distinguishes relevant from non-relevant segments (green line in Fig. 1). Speed-up rates are then calculated for each segment type, with more relevant segments assigned with lower rates. Each segment is represented as a graph where nodes correspond to frames and edges reflect transition costs between frames. The total edges cost includes terms that measure the inter-frame relevance drop, instability, motion speed, and appearance disparity, detailed by Halperin *et al.* [4] and Ramos *et al.* [9]. The frames in the shortest paths in these graphs are concatenated to compose the final video (Fig. 1).

TABLE I
QUANTITATIVE RESULTS. AVERAGE F_1 SCORE (%), SHAKING RATIO (%), AND OUTPUT SPEED-UP VALUES FOR THE OUTPUT VIDEOS.

Dataset	Method	F_1 Score ↑ CAR	CHAIR	COMP.	PEOPLE	TREE	Shaking Ratio ↓	Output Speed-up *
UTE	Unif.	09.6	**11.6**	10.8	12.2	10.2	31.1	-
	MSH	10.2	10.5	08.3	12.7	11.1	**27.0**	11.4
	MIFF	10.4	10.3	06.1	13.9	11.6	47.1	12.0
	Ours	**16.4**	10.1	**23.6**	**15.1**	**18.1**	37.2	**10.1**
Semantic Dataset	Unif.	12.9	07.3	06.9	08.1	15.2	11.0	-
	MSH	12.5	07.0	05.9	07.7	15.7	**04.4**	09.7
	MIFF	13.1	**09.1**	07.4	**13.6**	13.6	08.9	10.2
	Ours	**15.2**	08.8	**07.5**	12.4	**18.5**	10.1	**09.9**
Ego-Sequences	Unif.	12.8	03.7	02.2	15.4	17.9	12.0	-
	MSH	11.9	03.2	02.4	14.7	16.4	**04.7**	11.2
	MIFF	12.6	03.9	01.3	**17.2**	15.4	08.2	12.0
	Ours	**14.8**	**04.7**	**04.4**	16.4	**18.9**	08.2	**10.4**

**Better closer to 10.*

B. Experiments

Evaluation Setup. We used three datasets in our experiments: UT Egocentric (UTE) [21]; Semantic Dataset [10]; and EgoSequences [4], [9]. We compared our method against three first-person fast-forwarding approaches: *Uniform Sampling*, *Microsoft Hyperlapse* (MSH) [3], and *Multi-Importance Fast-Forward* (MIFF) [12]. For evaluation, we measured the personalization using the F_1 Score, the speed-up rate accuracy using the Output Speed-up (*i.e.*, $\hat{S} = |V|/|\hat{V}|$), and the instability using the Shaking Ratio, calculated as the average motion of the central point between frame transitions using homography transformations. In addition to real users, we created virtual users (character-based LSTMs) interested in common social network topics like Vehicles, Furniture, Technology, Human Interaction, and Nature for detailed evaluation. We used the parameters reported by Li *et al.* [36] for word embedding ($d = 300$) and the elbow method (from 2^1 to 2^{15}) to set $K = 2^{13}$. The target speed-up rate was set to $S^* = 10$.

Quantitative Results. We report the average F_1 Score, Shaking Ratio, and Output Speed-up values in Table I. Since only UTE contains human-annotated concepts, we used nouns from extracted sentences [34] to validate personalization in the other datasets. Our method generally yields higher personalization values across most concepts, especially in UTE, with an average F_1 Score 12.8 percentage points higher than the best competitor using tweets about COMPUTER. We accredit these results to our frame scoring approach, which effectively infers user topics and scores relevant scenes higher. Notable exceptions are the experiments with CHAIR, where Uniform and MIFF approaches outperformed ours in UTE and Semantic datasets by 1.5 and 0.3 percentage points, respectively. The reason is that this concept is constantly not the focus of

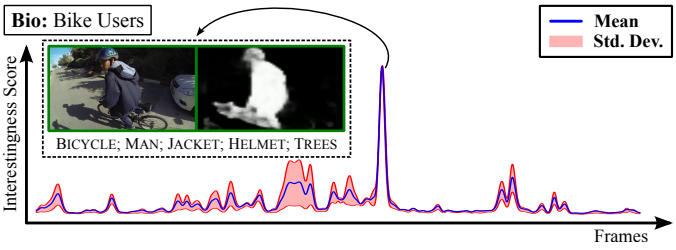

Fig. 2. **Scores for bike users in the video 'Walking 3' (EgoSequences).**

attention since it is always coupled with more prominent concepts like MEN. We argue that MIFF's better performance in the Semantic Dataset may be due to scenes featuring people seated in chairs or benches, which aligns with MIFF's focus on people. Regarding visual smoothness, MSH produced output videos with the lowest Shaking Ratio values, at least 3.5 percentage points better than the best competitor. This is an expected result since MSH directly optimizes the smoothness of its frame selection. All approaches presented Output Speed-up values close to the target, with our method achieving the best average values. Therefore, although our frame selection is more challenging than MSH's since it includes the semantics objective, it does not impair the speed-up rate accuracy.

Evaluation by Volunteers. We conducted a survey in which volunteers watched a video on a web page and were asked to (i) select the most emphasized content (shown at a lower speed-up rate) and (ii) evaluate the video's visual quality. Most volunteers (60.93% on average) selected the concept our method emphasizes. Notably, for the concept PEOPLE, our videos had more correct selections than MIFF (73.8% *vs.* 64.4%), despite MIFF focusing solely on people. Regarding visual quality, our method achieved an average score of 3 on a 5-point Likert Scale, similar to MIFF. This survey indicates that our method enhances semantic information encoding without compromising video stability.

Results with Twitter Users. We manually selected active public users on Twitter (five for each concept) who have indicated topics of interest in their profiles. We collected their last 3,000 tweets, when available, and applied our approach using representative videos from the datasets. Figure 2 shows the mean and standard deviation for the Interestingness Score assigned by our approach for cyclists on Twitter in the 'Walking 3' video from EgoSequences. The green box shows one of the frames with the highest score (left), along with the saliency map (right) and the extracted concepts (bottom). Despite cyclists' diverse interests in the video, the frame with high mutual interest features a man riding a bike, which is a unique moment in this video, highlighting the importance of the uniqueness score.

Limitations. Despite promising results, our method may fail to emphasize the relevant content if related concepts are spread across distinct clusters. Varying K or using a density-based clustering method can be an alternative. Additionally, a frame might receive a low score if its concept lacks visual saliency or has low TF-IDF due to its recurrence. Another drawback is the heavy reliance on off-the-shelf components, impacting both frame relevance and computational resources (time and memory). In the next section, we attempt to address this issue by creating an approach that directly models the input text and frames/segments. Thus, no intermediate steps like computing attention or extracting captions are required.

IV. FAST-FORWARDING INSTRUCTIONAL VIDEOS USING TEXTUAL INSTRUCTIONS

Instructional videos and online tutorials have been a key force in our modern educational routine. While both offer valuable information, they differ in consumption time. Textual descriptions are concise, summarizing tasks in a few sentences. In contrast, instructional videos often include extended, non-essential segments. Ideally, these videos should be concise yet visually demonstrate each key step effectively.

A. Methodology

We formulate the fast-forwarding task as a sequential decision-making process. An agent observes features of the encoded text and video frames, its position in the video, and its current average skip rate. Based on that, it decides to increase, decrease, or maintain the current speed-up rate.

Visually-guided Document Attention Network (VDAN). For the agent to align instructional text with video segments, we build an embedding space that encodes both documents and videos. Let v be a segment of length M from the input video $V = \{v_f\}_{f=1}^{F}$ of F frames, and $D = \{p_1, \cdots, p_N\}$ be a document composed of N sentences. We propose a Visually-guided Document Attention Network (VDAN) that produces embeddings $\mathbf{e}_f^v, \mathbf{e}_f^D \in \mathbb{R}^d$ to represent the visual and textual data, respectively (see Fig. 3-left).

We present VDAN in three variants: VDAN-S (Single-frame), VDAN-M (Multi-frames), and VDAN-T (Transformer-based). VDAN-S is the simpler architecture since it only encodes a single frame along with the document D, *i.e.* $M = 1$. For the visual branch, we use a ResNet-50 (pretrained on ImageNet) to extract features $\phi(v)$, projected into $\mathbb{R}^d$ by a fully connected network. In the textual branch, we employ a Hierarchical Recurrent Neural Network (H-RNN) with attention mechanisms at each level to weigh the importance of words and sentences. A key component in the VDAN architecture is the visual guidance, which bridges both modalities. To facilitate training and guide the attention weights during training, we set the first hidden vector of the H-RNN to $\phi(v)$. The output of the H-RNN is projected to $\mathbb{R}^d$ by another fully connected (FC) network. VDAN-M addresses temporal modeling by using an R(2+1)D-34 (pretrained on IG-65M) instead of ResNet-50, retaining the same remaining architecture as in VDAN-S. VDAN-T is based on Transformers, using the Space-Time Transformer Encoder (pretrained on WebVid-2M) for the visual encoder and BERT for the textual encoder at the sentence level. A cross-attention module, adapted from Multi-Head Attention, receives the sentences (to produce keys and values) and $\phi(v)$ to produce the query.

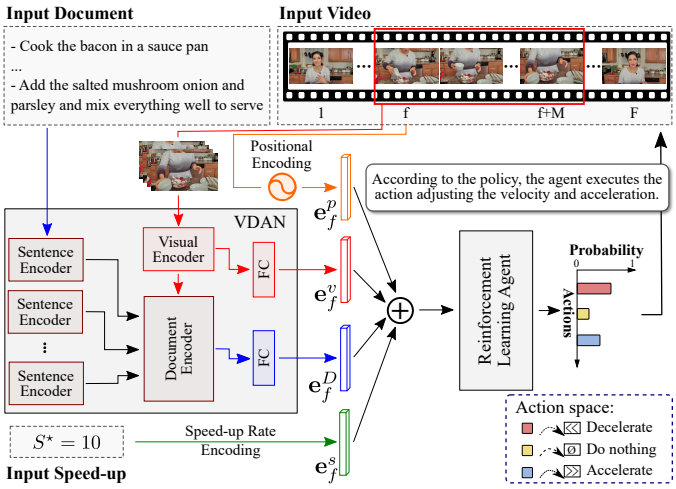

Fig. 3. **Proposed methodology.** Our approach involves two major training stages: *i)* creating a cross-modal embedding space (VDAN) to provide representative embeddings $\mathbf{e}_f^v$ for video segments and $\mathbf{e}_f^D$ for user documents; *ii)* training a reinforcement learning agent to select which frames to remove executing actions to increase, decrease, or keep the current skip rate given the embeddings $\mathbf{e}_f^v$ and $\mathbf{e}_f^D$, the encoded position in the video ($\mathbf{e}_f^p$), and the encoded speed-up rate ($\mathbf{e}_f^s$). The agent navigates throughout the video by skipping frames at each timestep according to the actions' probability.

We follow a pairwise training strategy to build the cross-modal embedding space. Video snippets (or still images) and their concatenated descriptions from human-annotated visual-textual datasets are used to compose positive and negative pairs. We apply the cosine embedding loss to the produced embeddings $\mathbf{e}_f^v$ and $\mathbf{e}_f^D$. To enhance the attention mechanism, we shuffle the positive sentences with sentences from a randomly sampled video, ensuring the attention mechanism correctly attends to the relevant sentences.

Semantic Fast-Forwarding via Reinforcement Learning (SFF-RL). We formulate the frame selection as a Markov Decision Process. We train an agent to observe the current semantics, its position in the video, and the average speed-up rate, then adjust the playback rate accordingly.

The agent's state at step t is composed of concatenated embeddings as: $\mathbf{s}_t = [\mathbf{e}_f^v; \mathbf{e}_f^D; \mathbf{e}_f^p; \mathbf{e}_f^s]^\intercal \in \mathcal{S}$, where $\mathbf{e}_f^D$ and $\mathbf{e}_f^v$ are from VDAN, and $\mathbf{e}_f^p$ and $\mathbf{e}_f^s$ are defined as follows. The $\mathbf{e}_f^p$ embedding encodes the location of the agent in the video; we name it Normalized Reversed Position Encoding (NRPE). Similar to positional encoding in Transformers, but reversed, it indicates how far the agent is from the video's end. It is normalized to ensure consistent actions when equally distant from different video ends. Let q be the NRPE embedding size. The dimensions $2k$ and $2k+1$ (with $k \in \{1, \cdots, q/2\}$) of our NRPE embedding are $NRPE_{(f,2k)} = \sin(F-f/F^{2k/q})$ and $NRPE_{(f,2k)} = \sin(F-f/F^{2k/q})$. The $\mathbf{e}_f^s$ encodes the agent's average skip rate; we name it Skip-Aware (SA). It is a one-hot vector defined as $\mathbf{e}_f^s = \mathbf{I}_m(\lfloor \hat{S}_t \rfloor - S^* + \nu_{max})$, where $\hat{S}_t$ is the average skip rate at the timestep t, $S^* \leq \nu_{max} \in \mathbb{N}^+$ is the target speed-up rate, $\nu_{max} = 25$ stands for the maximum skip rate the agent can achieve, and $\mathbf{I}_m(\iota)$ denotes the ι^{th} row of an identity matrix of size m.

We define the action space, $\mathcal{A}$, with three components: *decelerate*; *do nothing*; and *accelerate*. The *decelerate* action updates the agent's states as $\nu = \nu - \omega$ and $\omega = \omega - 1$, where ν is the current skip rate (velocity) and ω is the rate of change (acceleration). Similarly, the *accelerate* action updates the states as $\nu = \nu + \omega$ and $\omega = \omega + 1$. The *do nothing* action keeps ν and ω unchanged.

The goal of the agent is to maximize the expected sum of discounted rewards: $R_t = \mathbb{E}[\sum_{n=0}^{T-t} \gamma^n r_{t+n}]$, where t is the current timestep, r_{t+n} is the reward n timesteps into the future, T is the total number of timesteps, and $\gamma \in (0, 1]$ is a discount factor. Our immediate reward encourages the agent to adjust its skip rate based on the semantic similarity between the textual and visual data in the upcoming video segment while also considering the overall speed-up rate objective in the long term. At training time, after taking action $a_t \sim \pi(a|s_t, \theta_\pi)$ at step t, the agent receives the following reward signal:

$$r_t = \begin{cases} \mathbf{e}_f^D \cdot \mathbf{e}_f^v, & \text{if } t < T \\ \lambda * \exp(-0.5 * (\frac{\hat{S}_T - S^*}{\sigma})^2), & \text{otherwise,} \end{cases} \quad (1)$$

where λ controls the relative importance of the overall speed-up rate in relation to the frames' relevance in the output video. The terminal reward resembles a Gaussian function centered at S^*. Semantically, the agent receives higher rewards if $\mathbf{e}_f^D$ and $\mathbf{e}_f^v$ point in the same direction in the embedding space. This encourages the agent to reduce the speed and accumulate positive rewards, as neighboring temporal frames are likely to yield higher reward values due to their visual similarity.

We use REINFORCE [37] to train the policy $\pi(a|s_t, \theta_\pi)$ for the agent. Additionally, a state-value function $\delta(s_t|\theta_\delta)$ is trained with a regression loss to estimate R_t, reducing gradient variance. Both π and δ are fully-connected networks with parameters θ_π and θ_δ, respectively. At test time, the agent chooses the action $\arg\max_a \pi(a|s_t, \theta_\pi)$ at each timestep t.

B. Experiments

Evaluation Setup. We conducted our experiments in the YouCook2 [38] and COIN [39] datasets and compare our method against Sparse Adaptive Sampling (SAS) [13], SASv2 [14], Bag-of-Topics (BoT) from Sec. III, and Fast-Forward Network (FFNet) [8]. FFNet serves as a semantic benchmark since it disregards the target speed-up. Evaluation metrics include F_1 Score and Output Speed-up (OS). To balance these metrics, we introduce Overall Performance (OP), computed as the harmonic mean of F_1 and OS accuracy (the value in a Gaussian centered at S^*). For VDAN, we set $d = 128$ and train it for 100 epochs in VDAN-S using MSCOCO [40] and for 30 epochs in VDAN-M/T using VaTeX [32] with learning rate of $1e{-}5$. For SFF-RL, we freeze the VDAN weights and train the agent for 100 epochs using the Adam optimizer with a learning rate of $5e{-}5$ for the policy and $1e{-}3$ for the state-value. After running a grid search, we set $\sigma = 0.5$ and $\gamma = 0.99$. We empirically set $\lambda = F$ for VDAN-S and $\lambda = F^*$ for VDAN-M/T, where F^* is the target number of frames.

Quantitative Results. Table II shows the results for all

TABLE II
COMPARISON WITH BASELINES. THE TARGET OS VALUES ARE $S^* = 12$ FOR YOUCOOK2 AND $S^* = 16$ FOR COIN. BEST AND SECOND-BEST VALUES ARE IN BOLD AND ITALICS, RESPECTIVELY.

		Baselines			Ours			
		SAS	SASv2	BoT	VDAN-S	VDAN-M	VDAN-T	FFNet
YouCook2	F_1[1]	14.44	16.20	14.05	14.36	**17.86**	*17.49*	18.86
	OS[2]	11.64	10.32	**12.14**	*12.24*	11.68	11.75	11.90
	OP[1]	25.01	19.06	24.61	25.03	**30.07**	*29.63*	31.72
COIN	F_1[1]	13.20	13.90	13.01	13.40	**17.18**	*14.73*	17.66
	OS[2]	*16.10*	14.09	**16.07**	16.67	14.99	16.22	16.45
	OP[1]	22.82	19.78	23.02	23.24	**27.98**	*25.64*	29.74

[1]*Higher is better (%)* [2]*Better closer to* S^*

compared methods in both datasets. Since FFNet does not allow a target speed-up rate, we used its average OS, $\lfloor \hat{S} \rceil = 12$ (YouCook2) and $\lfloor \hat{S} \rceil = 16$ (COIN), as targets for all methods.

The results show that our approach in at least one of its VDAN variants significantly outperforms all competitors in F_1, as verified by a t-test, without significantly compromising the output speed-up in both datasets, as reflected in the OP metric. It means that our agents effectively use natural language instructions to match them to the current scene and decide what to skip. Conversely, SAS and SASv2, which use YOLO for semantic encoding, did not perform well due to a lack of detail in object interactions requiring motion modeling. BoT also underperformed, likely due to its reliance on the accuracy of external components. For instance, in many cases, a person carrying out the task is given higher saliency than the ingredients themselves.

Qualitative Results. Fig. 4 shows a frame selection performed by our agent using VDAN-M as the semantic encoder. The green vertical bars represent the selected frames, and the black contiguous blocks the ground truth. Note that our method performs a denser frame selection on segments where the instruction is depicted (images 1 and 3). The agent increases the video playback speed, producing a sparser frame selection, mostly in segments with no instructions (image 2). Occasionally, the agent may erroneously reduce speed in non-instruction regions due to high visual-text similarity (image 4) or even skip relevant frames to meet the desired speed-up rate.

Ablation Studies. We tested our agent's ability to meet target speed-up rates by running the inference using rates from 2 to 20 and the same trained agent. We verified that the absolute errors (*i.e.*, $|\hat{S} - S^*|$) are, on average, 0.55, indicating effective control of our agent over the output video's length.

We evaluated the impact of each component in our approach using the VDAN-M variant in YouCook2. The SA element is crucial for meeting the target speed-up rate; without it, the agent either accelerates the entire video at ν_{max} or does not accelerate at all. Without the NRPE component, the agent is cautious about skip rates, avoiding significant speed changes to preserve frame relevance, as it lacks awareness of the video's end. The agent with all proposed components balances relevance and speed-up rate best. We also tried a non-RL variant by replacing our agent with the BoT frame selector and using the $\mathbf{e}_f^D$ and $\mathbf{e}_f^v$ similarity scores. Results confirm the superiority of our SFF-RL stage not only in the F_1, OS, and OP metrics but also in running time, taking 13.87× less.

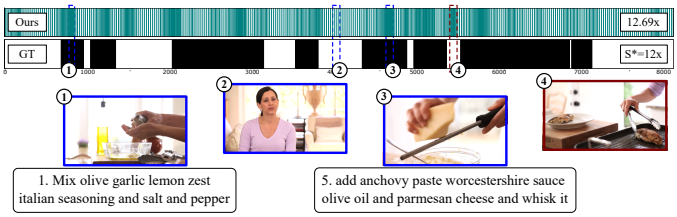

Fig. 4. **Our method's frame selection (VDAN-M).** Green vertical bars represent the selected frames and black contiguous blocks the ground-truth. Our method performs a denser frame selection on relevant segments (images 1 and 3) while discarding irrelevant ones (image 2) to meet the user's desired speed-up rate ($\hat{S} = 12.69$). The agent may erroneously select frames in regions without instructions (image 4) due to their similarity to the input text.

Limitations. Despite achieving the best results, our methodology has limitations. The agent may incorrectly emphasize non-instructional segments (see Fig. 4, image 4). The reward function is sparse regarding speed-up rate deviation and intensifies as the agent reaches the video's end, causing it to sometimes disregard relevant segments near the end.

V. CONCLUSION

In this work, we tackled the problem of fully automatic acceleration using web texts as semantic clues to define the frames' relevance for the user. We approached this problem in two different contexts, developing methodologies capable of extracting semantic information from natural language to identify visual concepts and events of higher relevance. Quantitative and qualitative experiments, ablation studies, and a user study demonstrate the superiority of our methods over the baselines. Although the second method takes a significant step towards an end-to-end approach, reducing the reliance on off-the-shelf components, some challenges remain, particularly in avoiding the emphasis on non-relevant segments. Future directions include using audio as supervision, which could help disambiguate relevance and reduce the need for human annotations since it is commonly present and naturally aligned with the visual stream. Another direction is employing multiple agents for video acceleration to collaboratively gain full video context and overcome limitations like sparse rewards.

Acknowledgments. We thank CAPES, CNPq, FAPEMIG, and Petrobras for funding different parts of this work.

VI. AWARDS & PUBLICATIONS

This work was awarded with the Microsoft Research PhD Fellowship program in 2021. Part of this work was published in the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) in 2023 [41], at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [42], and at the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) [43].

REFERENCES

[1] J. Kopf, M. F. Cohen, and R. Szeliski, "First-person hyper-lapse videos," Proc. of the ACM Trans. on Graph., vol. 33, no. 4, 2014.

[2] Y. Poleg, T. Halperin, C. Arora, and S. Peleg, "EgoSampling: Fast-forward and stereo for egocentric videos," in Proc. of the IEEE Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2015, pp. 4768–4776.

[3] N. Joshi, W. Kienzle, M. Toelle, M. Uyttendaele, and M. F. Cohen, "Real-time hyperlapse creation via optimal frame selection," Proc. of the ACM Trans. on Graph., vol. 34, no. 4, 2015.

[4] T. Halperin, Y. Poleg, C. Arora, and S. Peleg, "EgoSampling: Wide view hyperlapse from egocentric videos," IEEE Trans. on Circuits and Sys. for Video Technology, vol. 28, no. 5, pp. 1248–1259, 2018.

[5] P. Rani, A. Jangid, V. P. Namboodiri, and K. S. Venkatesh, "Visual odometry based omni-directional hyperlapse," in Proc. of the National Conf. on Comp. Vis., Patt. Rec., Image Proc., and Graph., Singapore, 2018, pp. 3–13.

[6] M. Wang, J. Liang, S. Zhang, S. Lu, A. Shamir, and S. Hu, "Hyper-lapse from multiple spatially-overlapping videos," IEEE Trans. on Image Proc., vol. 27, no. 4, pp. 1735–1747, 2018.

[7] M. Okamoto and K. Yanai, "Summarization of egocentric moving videos for generating walking route guidance," in Proc. of the Pacific-Rim Symposium on Image and Video Technology, 2013, pp. 431–442.

[8] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury, "FFNet: Video fast-forwarding via reinforcement learning," in Proc. of the IEEE/CVF Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2018, pp. 6771–6780.

[9] W. L. S. Ramos, M. M. Silva, M. F. M. Campos, and E. R. Nascimento, "Fast-forward video based on semantic extraction," in Proc. of the IEEE Int. Conf. on Image Proc. (ICIP), 2016, pp. 3334–3338.

[10] M. M. Silva, W. L. S. Ramos, J. P. K. Ferreira, M. F. M. Campos, and E. R. Nascimento, "Towards semantic fast-forward and stabilized egocentric videos," in Proc. of the Eur. Conf. on Comp. Vis. Workshop (ECCVW), 2016, pp. 557–571.

[11] W.-S. Lai, Y. Huang, N. Joshi, C. Buehler, M.-H. Yang, and S. B. Kang, "Semantic-driven generation of hyperlapse from 360° video," IEEE Trans. on Vis. and Com. Graph., vol. 24, no. 9, pp. 2610–2621, 2017.

[12] M. M. Silva, W. L. Ramos, F. C. Chamone, J. P. Ferreira, M. F. Campos, and E. R. Nascimento, "Making a long story short: A multi-importance fast-forwarding egocentric videos with the emphasis on relevant objects," Journal of Vis. Comm. and Image Rep., vol. 53, pp. 55–64, 2018.

[13] M. Silva, W. Ramos, J. Ferreira, F. Chamone, M. Campos, and E. R. Nascimento, "A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos," in Proc. of the IEEE/CVF Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2018, pp. 2383–2392.

[14] M. Silva, W. Ramos, M. Campos, and E. R. Nascimento, "A sparse sampling-based framework for semantic fast-forward of first-person videos," IEEE Trans. on Patt. Anal. and Mach. Intel. (TPAMI), vol. 43, no. 4, pp. 1438–1444, 2021.

[15] S. Lan, Z. Wang, A. K. Roy-Chowdhury, E. Wei, and Q. Zhu, "Distributed multi-agent video fast-forwarding," in Proc. of the 28th ACM Int. Conf. on Multim., ser. MM '20, New York, NY, USA, 2020, pp. 1075–1084.

[16] D. de Matos, W. Ramos, L. Romanhol, and E. R. Nascimento, "Musical hyperlapse: A multimodal approach to accelerate first-person videos," in 2021 34th SIBGRAPI Conf. on Graphics, Patt. and Images (SIBGRAPI), 2021, pp. 184–191.

[17] S. Lan, Z. Wang, E. Wei, A. K. Roy-Chowdhury, and Q. Zhu, "Collaborative multi-agent video fast-forwarding," IEEE Trans. on Multim., pp. 1–14, 2023.

[18] D. de Matos, W. Ramos, M. Silva, L. Romanhol, and E. R. Nascimento, "A multimodal hyperlapse method based on video and songs' emotion alignment," Patt. Rec. Letters, vol. 166, pp. 174–181, 2023.

[19] H. Shvaytser, "Learnable and nonlearnable visual concepts," IEEE Trans. on Patt. Anal. and Mach. Intel. (TPAMI), vol. 12, no. 5, pp. 459–466, 1990.

[20] T. Li and L. Wang, "Learning spatiotemporal features via video and text pair discrimination," CoRR, vol. abs/2001.05691, 2021.

[21] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in Proc. of the IEEE Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2012, pp. 1346–1353.

[22] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in Proc. of the AAAI Conf. on Artif. Intel., 2018, pp. 7582–7589.

[23] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in Proc. of the IEEE Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2016, pp. 982–990.

[24] P. Varini, G. Serra, and R. Cucchiara, "Personalized egocentric video summarization of cultural tour on user preferences input," IEEE Trans. on Multimedia, vol. 19, no. 12, pp. 2832–2845, 2017.

[25] A. Sharghi, J. S. Laurel, and B. Gong, "Query-focused video summarization: Dataset, evaluation, and a memory network based approach," in Proc. of the IEEE Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2017, pp. 4788–4797.

[26] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in Proc. of the IEEE/CVF Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2019, pp. 7894–7903.

[27] Z. Li and L. Yang, "Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward," in IEEE Winter Conf. on App. of Comp. Vis. (WACV), 2021, pp. 3239–3247.

[28] H. Jiang and Y. Mu, "Joint video summarization and moment localization by cross-task sample transfer," in Proc. of the IEEE/CVF Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2022, pp. 16 388–16 398.

[29] K. Higuchi, R. Yonetani, and Y. Sato, "EgoScanning: Quickly scanning first-person videos with egocentric elastic timelines," in Proc. of the Conf. on Human Factors in Computing Sys. (CHI), ser. CHI '17, New York, NY, USA, 2017, pp. 6536–6546.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry et al., "Learning transferable visual models from natural language supervision," Image, vol. 2, p. T2, 2021.

[31] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in Proc. of the IEEE/CVF Int. Conf. on Comp. Vis. (ICCV), 2021, pp. 1728–1738.

[32] X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Y. Wang, "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research," in Proc. of the IEEE/CVF Int. Conf. on Comp. Vis. (ICCV), 2019, pp. 4580–4590.

[33] X. Jiang, X. Xu, J. Zhang, F. Shen, Z. Cao, and H. T. Shen, "Semi-supervised video paragraph grounding with contrastive encoder," in Proc. of the IEEE/CVF Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2022, pp. 2466–2475.

[34] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in Proc. of the IEEE Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2016, pp. 4565–4574.

[35] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," IEEE Trans. on Image Proc., vol. 27, no. 1, pp. 38–49, 2018.

[36] Q. Li, S. Shah, X. Liu, and A. Nourbakhsh, "Data sets: Word embeddings learned from tweets and general data," CoRR, vol. abs/1708.03994, 2017.

[37] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," Machine Learning, vol. 8, no. 3, pp. 229–256, 1992.

[38] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in Proc. of the AAAI Conf. on Artif. Intel., 2018, pp. 7590–7598.

[39] Y. Tang, J. Lu, and J. Zhou, "Comprehensive instructional video analysis: The coin dataset and performance evaluation," IEEE Trans. on Patt. Anal. and Mach. Intel. (TPAMI), pp. 1–1, 2020.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Proc. of the Eur. Conf. on Comp. Vis. (ECCV), 2014, pp. 740–755.

[41] W. Ramos, M. Silva, E. Araujo, V. Moura, K. Oliveira, L. S. Marcolino, and E. R. Nascimento, "Text-driven video acceleration: A weakly-supervised reinforcement learning method," IEEE Trans. on Patt. Anal. and Mach. Intel. (TPAMI), vol. 45, no. 2, pp. 2492–2504, 2023.

[42] W. Ramos, M. Silva, E. Araujo, L. S. Marcolino, and E. Nascimento, "Straight to the point: Fast-forwarding videos via reinforcement learning using textual data," in Proc. of the IEEE/CVF Conf. on Comp. Vis. and Patt. Rec. (CVPR), 2020, pp. 10 928–10 937.

[43] W. L. S. Ramos, M. M. Silva, E. R. Araujo, A. C. Neves, and E. R. Nascimento, "Personalizing fast-forward videos based on visual and textual features from social network," in IEEE Winter Conf. on App. of Comp. Vis. (WACV), 2020, pp. 3260–3269.