Causalainer: Causal Explainer for Automatic Video Summarization

Jia-Hong Huang[†]
University of Amsterdam
j.huang@uva.nl

Chao-Han Huck Yang
Georgia Institute of Technology
huckiyang@gatech.edu

Pin-Yu Chen
IBM Research AI
pin-yu.chen@ibm.com

Min-Hung Chen
NVIDIA (ex-Microsoft)
minhungc@nvidia.com

Marcel Worring
University of Amsterdam
m.worring@uva.nl

Abstract

The goal of video summarization is to automatically shorten videos such that it conveys the overall story without losing relevant information. In many application scenarios, improper video summarization can have a large impact. For example in forensics, the quality of the generated video summary will affect an investigator's judgment while in journalism it might yield undesired bias. Because of this, modeling explainability is a key concern. One of the best ways to address the explainability challenge is to uncover the causal relations that steer the process and lead to the result. Current machine learning-based video summarization algorithms learn optimal parameters but do not uncover causal relationships. Hence, they suffer from a relative lack of explainability. In this work, a Causal Explainer, dubbed Causalainer, is proposed to address this issue. Multiple meaningful random variables and their joint distributions are introduced to characterize the behaviors of key components in the problem of video summarization. In addition, helper distributions are introduced to enhance the effectiveness of model training. In visual-textual input scenarios, the extra input can decrease the model performance. A causal semantics extractor is designed to tackle this issue by effectively distilling the mutual information from the visual and textual inputs. Experimental results on commonly used benchmarks demonstrate that the proposed method achieves state-of-the-art performance while being more explainable.

1. Introduction

Video summarization is the process of automatically generating a concise video clip that conveys the primary message or story in the original video. Various automatic video summarization algorithms have been proposed in recent years to tackle this task using different supervision schemes. These include fully-supervised methods that utilize visual input alone [10, 13, 35, 36, 68–71] or multi-modal input [26, 27, 42, 43, 46, 51, 54, 56, 58, 65, 72], as well as weakly-supervised methods [4, 6, 17, 38, 47, 60].

[†]Work done during an internship at Microsoft Research in Cambridge, UK and Amsterdam, NL.

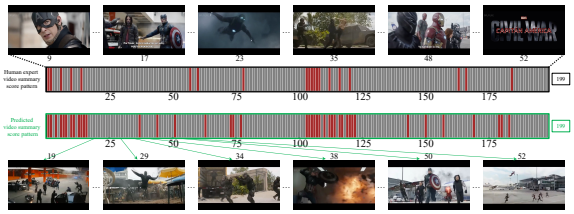

Figure 1. Visualization of human-annotated and machine-predicted frame-level scores for creating a video summary. Comparing the human-annotated video summary score pattern to the one generated by existing state-of-the-art video summarization methods, e.g., [26, 27, 56], we observe these methods are capable of learning visual consecutiveness and diversity which are some key factors considered by humans for creating a good video summary. These methods mainly focus on capturing the visual cues to achieve such a purpose. Red bars denote discarded frames and grey bars indicate selected frames used to form a summary. The video has 199 frames and the numbers, except for 199, denote the indices of frames.

According to [10, 13, 14, 27, 55, 56], when human experts perform the task of video summary generation, they will not only consider concrete/visual factors, e.g., visual consecutiveness and visual diversity, but also abstract/non-visual factors, such as interestingness, representativeness, and storyline smoothness. Hence, a human-generated video summary is based on many confounding factors. These factors/causes result in the video summary. Existing works do not, or in a very limited way, consider abstract factors and mainly focus on proposing various methods to exploit concrete visual cues to perform video summarization. See the illustration in Figure 1. This leads to limited modeling explainability of automatic video summarization [27, 37].

Machine learning (ML) models can be made more explainable through causation modeling based on Bayesian probability [39, 48, 49, 66, 67]. In this work, we propose a novel method for improving the inherent explainability of video summarization models called Causalainer, which is based on causation modeling. Our approach aims to address the challenge of model explainability in video summarization by leveraging the insights gained from Bayesian probability and causation modeling. See Figure 2 for the method flowchart of the proposed Causalainer. To model the problem of video summarization and increase the explainability, four meaningful random variables are introduced to characterize the behaviors of the data intervention, the model's

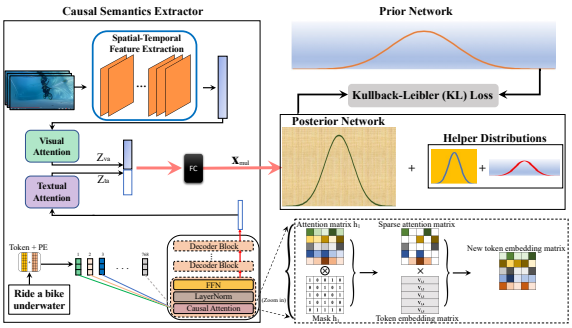

Figure 2. Flowchart of the proposed Causal Explainer (Causalainer) method for video summarization. The proposed method is mainly composed of a prior network, a posterior network, helper distributions, and a causal semantics extractor. $\otimes$ denotes element-wise multiplication and $\times$ indicates matrix multiplication. "Token + PE" denotes the operations of token embedding and positional encoding.

prediction, observed potential confounders, and unobserved confounders, respectively. Note that data intervention is a way to help a model learn the causal relations that lead to the result [5, 11, 12, 45, 52, 53]. A prior joint distribution and its posterior approximation can be built on top of those four random variables. The proposed method is trained based on minimizing the distance between the prior distribution and the posterior approximation. We identify that predicting the behaviors of the data intervention and model's outcome can be challenging in practice due to various factors, e.g., video noise, lens or motion blur. We address this issue by introducing helper distributions for them. The helper distributions form a new loss term to guide the model learning. Furthermore, when multi-modal inputs are available, we identify that the extra input sometimes can harm the model performance most likely due to the interactions between different modalities being ineffective. We address this challenge by introducing a causal semantics extractor to effectively distill the mutual information between multi-modal inputs.

These novel design choices have been instrumental in improving the explainability and performance of video summarization models. The extensive experimentation on commonly used video summarization datasets verifies that the proposed method outperforms existing state-of-the-art while also providing greater explainability. By leveraging causal learning techniques, our approach represents a promising attempt to reinforce the causal inference ability and explainability of an ML-based video summarization model.

2. Methodology

We now present the details of the proposed Causal Explainer method for automatic video summarization, dubbed Causalainer. First, the assumptions of causal modeling are described in detail. Secondly, we introduce four random variables $\mathbf{y}$, $\mathbf{t}$, $\mathbf{X}$, and $\mathbf{Z}$ to characterize the behaviors of the model's prediction, the data intervention, observed potential confounders, and unobserved confounders, respectively. Finally, the derivation of our training objective with helper distributions and the proposed causal semantics extractor are presented. Causalainer consists of prior and posterior probabilistic networks. See Figure 2 for an overview.

3.1 Assumptions

In general, causal learning for real-world observational studies is complicated [1, 2, 7, 19–25, 28–34, 44, 45, 59, 61–63]. With the established efforts [45, 49, 66] on causal learning under noisy interventions, two assumptions are imposed when modeling the problem of video summarization. First, the information of having visual/textual intervention $\mathbf{t}$ or not is binary. Second, the observations $(\mathbf{X}, \mathbf{t}, \mathbf{y})$ from a deep neural network (DNN) are sufficient to approximately recover the joint distribution $p(\mathbf{Z}, \mathbf{X}, \mathbf{t}, \mathbf{y})$ of the unobserved/latent confounding variable $\mathbf{Z}$, the observed confounding variable $\mathbf{X}$, the intervention $\mathbf{t}$, and the outcome $\mathbf{y}$. The proposed Causalainer method is built on top of multiple probability distributions as described in the following subsections.

3.2 Causal Explainer for Video Summarization

In the proposed Causalainer, $\mathbf{x}_i$ denotes an input video and an optional text-based query indexed by i, $\mathbf{z}_i$ indicates the latent confounder, $t_i \in \{0, 1\}$ denotes the intervention assignment, and y_i indicates the outcome.

Prior Probability Distributions. The prior network is conditioning on the latent variable $\mathbf{z}_i$ and mainly consists of the following components: **(i)** The latent confounder distribution: $p(\mathbf{z}_i) = \prod_{z \in \mathbf{z}_i} \mathcal{N}(z|\mu = 0, \sigma^2 = 1)$, where $\mathcal{N}(z|\mu, \sigma^2)$ denotes a Gaussian distribution with a random variable z, z is an element of $\mathbf{z}_i$, and the mean μ and variance σ^2 follow the settings in [41], i.e., $\mu = 0$ $\sigma^2 = 1$. **(ii)** The conditional data distribution: $p(\mathbf{x}_i|\mathbf{z}_i) = \prod_{x \in \mathbf{x}_i} p(x|\mathbf{z}_i)$, where $p(x|\mathbf{z}_i)$ is an appropriate probability distribution with a random variable x, the distribution is conditioning on $\mathbf{z}_i$, and x is an element of $\mathbf{x}_i$. **(iii)** The conditional intervention distribution: $p(t_i|\mathbf{z}_i) = \text{Bernoulli}(\sigma(f_{\theta_1}(\mathbf{z}_i)))$, where $\sigma(\cdot)$ is a logistic function, $\text{Bernoulli}(\cdot)$ indicates a Bernoulli distribution for a discrete outcome, and $f_{\theta_1}(\cdot)$ denotes a neural network parameterized by the parameter θ_1. **(iv)** The conditional outcome distribution: $p(y_i|\mathbf{z}_i, t_i) = \sigma(t_i f_{\theta_2}(\mathbf{z}_i) + (1 - t_i) f_{\theta_3}(\mathbf{z}_i))$, where $f_{\theta_2}(\cdot)$ and $f_{\theta_3}(\cdot)$ are neural networks parameterized by the parameters θ_2 and θ_3, respectively. In this work, y_i is tailored for a categorical classification problem, i.e., frame-based importance score classification in video summarization.

Posterior Probability Distribution. Since a priori knowledge on the latent confounder does not exist, we have to marginalize over it in order to learn the model parameters, θ_1, θ_2, and θ_3 in **(iii)** and **(iv)**. The non-linear neural network functions make inference intractable. Hence, variational inference [41] along with the posterior network is employed. These neural networks output the parameters of a fixed form posterior approximation over the latent variable $\mathbf{z}$, given the observed variables. Similar to [45, 50], in this work, the proposed posterior network is conditioning on observations. Also, the true posterior over $\mathbf{Z}$ depends on $\mathbf{X}$, $\mathbf{t}$ and $\mathbf{y}$. Hence, the posterior approximation defined below is employed to build the posterior network. $q(\mathbf{z}_i|\mathbf{x}_i, y_i, t_i) = \prod_{z \in \mathbf{z}_i} \mathcal{N}(z|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$, where $\boldsymbol{\mu}_i = t_i \boldsymbol{\mu}_{t=1,i} + (1 - t_i)\boldsymbol{\mu}_{t=0,i}$, $\boldsymbol{\sigma}_i^2 = t_i \boldsymbol{\sigma}_{t=1,i}^2 + (1 - t_i)\boldsymbol{\sigma}_{t=0,i}^2$, $\boldsymbol{\mu}_{t=0,i} = g_{\phi_1} \circ g_{\phi_0}(\mathbf{x}_i, y_i)$, $\boldsymbol{\sigma}_{t=0,i}^2 = \sigma(g_{\phi_2} \circ g_{\phi_0}(\mathbf{x}_i, y_i))$, $\boldsymbol{\mu}_{t=1,i} = g_{\phi_3} \circ g_{\phi_0}(\mathbf{x}_i, y_i)$, $\boldsymbol{\sigma}_{t=1,i}^2 = \sigma(g_{\phi_4} \circ g_{\phi_0}(\mathbf{x}_i, y_i))$, $g_{\phi_k}(\cdot)$ denotes a neural network with variational parameters ϕ_k for $k = 0, 1, 2, 3, 4$, and $g_{\phi_0}(\mathbf{x}_i, y_i)$ is a shared representation. Note that a feature map is multiplied with the approximated posterior $q(y_i|\mathbf{x}_i, t_i)$ without logistic function $\sigma(\cdot)$ to get $g_{\phi_0}(\mathbf{x}_i, y_i)$.

3.3 Training Objective with Helper Distributions

In practice, various factors, e.g., video noise, motion blur, or lens blur, make the prediction of the behaviors of the data intervention and the model's outcome challenging. Therefore, two helper distributions are introduced to alleviate this issue. We have to know the intervention assignment $\mathbf{t}$ along with its outcome $\mathbf{y}$ before inferring the distribution over $\mathbf{Z}$. Hence, the helper distribution $q(t_i|\mathbf{x}_i) = \text{Bernoulli}(\sigma(g_{\phi_5}(\mathbf{x}_i)))$ is introduced for the intervention assignment t_i, and the other helper distribution $q(y_i|\mathbf{x}_i, t_i) = \sigma(t_i g_{\phi_6}(\mathbf{x}_i) + (1-t_i) g_{\phi_7}(\mathbf{x}_i))$ is introduced for the outcome y_i, where $g_{\phi_k}(\cdot)$ indicates a neural network with variational parameters ϕ_k for $k = 5, 6, 7$. The introduced helper distributions benefit the prediction of t_i and y_i for new samples. To estimate the variational parameters of the distributions $q(t_i|\mathbf{x}_i)$ and $q(y_i|\mathbf{x}_i, t_i)$, a helper objective function $\mathcal{L}_{\text{helper}} = \sum_{i=1}^{N} [\log q(t_i = t_i^*|\mathbf{x}_i^*) + \log q(y_i = y_i^*|\mathbf{x}_i^*, t_i^*)]$ is introduced to the final training objective over N data samples, where $\mathbf{x}_i^*$, t_i^* and y_i^* are the observed values in the training set. The overall training objective $\mathcal{L}_{\text{causal}}$ for the proposed method is defined below. $\mathcal{L}_{\text{causal}} = \mathcal{L}_{\text{helper}} + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i, t_i, y_i)} [\log p(\mathbf{x}_i, t_i|\mathbf{z}_i) + \log p(y_i|t_i, \mathbf{z}_i) + \log p(\mathbf{z}_i) - \log q(\mathbf{z}_i|\mathbf{x}_i, t_i, y_i)]$.

3.4 Causal Semantics Extractor

Existing commonly used video summarization datasets, e.g., TVSum [55] and QueryVS [27], provide visual and textual inputs. Since the textual input cannot always help the model performance because of the ineffective extraction of mutual information from the visual and textual inputs, a causal semantics extractor is introduced to alleviate this issue. The proposed extractor is built on top of transformer blocks [57]. Vanilla transformers exploit all of the tokens in each layer for attention computation. However, the design philosophy of the proposed causal semantics extractor, dubbed causal attention, is effectively using fewer but relatively informative tokens to compute attention maps, instead of using the total number of tokens. According to [57], the computation of the vanilla attention matrix $\mathscr{A} \in \mathbb{R}^{n \times n}$ is based on the dot-product. It is defined as $\mathscr{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right); \mathbf{Q} = \mathbf{T}\mathbf{W}_q, \mathbf{K} = \mathbf{T}\mathbf{W}_k$, where the query matrix $\mathbf{Q} \in \mathbb{R}^{n \times d}$ and key matrix $\mathbf{K} \in \mathbb{R}^{n \times d}$ are generated by the linear projection of the input token matrix $\mathbf{T} \in \mathbb{R}^{n \times d_m}$ based on the learnable weights matrices $\mathbf{W}_q \in \mathbb{R}^{d_m \times d}$ and $\mathbf{W}_k \in \mathbb{R}^{d_m \times d}$. n indicates the total number of input tokens. d represents the embedding dimension and d_m denotes the dimension of an input token. The new value matrix $\mathbf{V}_{\text{new}} \in \mathbb{R}^{n \times d}$ can be obtained via $\mathbf{V}_{\text{new}} = \mathscr{A}\mathbf{V}; \mathbf{V} = \mathbf{T}\mathbf{W}_v$, where the value matrix $\mathbf{V} \in \mathbb{R}^{n \times d}$ and $\mathbf{W}_v \in \mathbb{R}^{d_m \times d}$.

In [57], the vanilla attention matrix is based on the calculation of all the query-key pairs. However, in the proposed Causal Semantics Extractor, only the top κ most similar keys and values for each query are used to compute the causal attention matrix. Similar to [57], all the queries and keys are calculated by the dot-product. Then, the row-wise top κ elements are used for the softmax calculation. In the proposed Causal Semantics Extractor, the value matrix $\mathbf{V}_\kappa \in \mathbb{R}^{n \times d}$ is defined as $\mathbf{V}_\kappa = \text{softmax}\left(\tau_\kappa(\mathscr{A})\right)\mathbf{V}_{\text{new}} = \text{softmax}\left(\tau_\kappa\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\right)\mathbf{V}_{\text{new}}$, where $\tau_\kappa(\cdot)$ denotes an operator for the row-wise top κ elements selection. $\tau_\kappa(\cdot)$ is defined as $[\tau_\kappa(\mathscr{A})]_{ij} = \begin{cases} \mathscr{A}_{ij} & , \mathscr{A}_{ij} \in \text{top } \kappa \text{ factors at row } i \\ -\infty & , \text{otherwise.} \end{cases}$

Then, $\mathbf{V}_\kappa$ can be further used to generate $\mathbf{X}_{\text{mul}}$, i.e., an output of the proposed Causal Semantics Extractor. The procedure for calculating $\mathbf{X}_{\text{mul}}$ is defined below. $Z_{\text{ta}} = \text{TextAtten}(\text{FFN}(\text{LayerNorm}(\mathbf{V}_\kappa))$, where $\text{LayerNorm}(\cdot)$ denotes a layer normalization, $\text{FFN}(\cdot)$ indicates a feed forward network, and $\text{TextAtten}(\cdot)$ denotes an element-wise multiplication-based textual attention mechanism. $Z_{\text{va}} = \text{VisualAtten}(\text{C3D}(\mathbf{I}))$, where I denotes an input video, $\text{C3D}(\cdot)$ indicates an operation of the spatial-temporal feature extraction, e.g., 3D version of ResNet-34 [15, 16], for the input video, and $\text{VisualAtten}(\cdot)$ indicates a visual attention mechanism based on the element-wise multiplication. $\mathbf{X}_{\text{mul}} = \text{FC}(Z_{\text{ta}} \odot Z_{\text{va}})$, where $\odot$ denotes the operation of feature concatenation and $\text{FC}(\cdot)$ indicates a fully connected layer. Note that the Causal Semantics Extractor's output $\mathbf{X}_{\text{mul}}$ is an input of the proposed posterior network based on the scheme of using multi-modal inputs.

Similar to the final step of video summary generation in [27], after the end-to-end training of the proposed causal video summarization model is complete, the trained model can be used for video summary generation. Finally, based on the generated score labels, a set of video frames is selected from the original input video to form a final video summary. Note that the summary budget is considered as a user-defined hyper-parameter in multi-modal video summarization [27].

3. Experiments

3.1 Experimental Setup and Datasets Preparation

Experimental Setup. We consider three scenarios: 1) fully-supervised training with human-defined frame-level labels, 2) fully-supervised training with multi-modal input including text-based query, and 3) weakly-supervised learning with two-second segment-level scores, which can be considered as a form of weak label [3, 4, 6]. Note that [55] empirically finds that a two-second segment length is appropriate for capturing video local context with good visual coherence. Hence, in this work, a video segment-level score is produced per two seconds based on given frame-level scores.

Video Summarization Datasets. In the experiments, three commonly used video summarization datasets, i.e., TVSum [55], QueryVS [27], and SumMe [13], are exploited to evaluate the proposed method. The TVSum dataset contains 50 videos. The length of the video in TVSum is ranging from 2 to 10 minutes. The human expert frame-level importance score label in TVSum is ranging from 1 to 5. The QueryVS dataset contains 190 videos. The video length in QueryVS is ranging from 2 to 3 minutes. The human expert frame-level importance score label in QueryVS is ranging from 0 to 3. Every video is retrieved based on a given text-based query. The SumMe dataset contains 25 videos. The video duration in SumMe is ranging from 1 to 6 minutes. In SumMe, the importance score annotated by human experts ranges from 0 to 1. Note that SumMe is not used for multi-modal video summarization. Hence, we do not have textual input when a model is evaluated on this dataset. Videos from these datasets are sampled at 1 frame per second (fps). The input image size is 224 by

Table 1. Comparison with fully-supervised state-of-the-art methods. The proposed method performs the best on both datasets. Note that textual query input is not used in this experiment.

Fully-supervised Method	TVSum	SumMe
SASUM [58]	53.9	40.6
dppLSTM [68]	54.7	38.6
ActionRanking [8]	56.3	40.1
H-RNN [70]	57.7	41.1
CRSum [64]	58.0	47.3
M-AVS [36]	61.0	44.4
VASNet [9]	61.4	49.7
iPTNet [37]	63.4	54.5
DASP [35]	63.6	45.5
Causalainer	**67.5**	**52.4**

Table 2. Comparison with the multi-modal state-of-the-art. The proposed method outperforms the existing multi-modal approaches. '-' denotes unavailability from previous work.

Multi-modal Method	TVSum	QueryVS
DSSE [65]	57.0	-
QueryVS [27]	-	41.4
DQSN [72]	58.6	-
GPT2MVS [26]	-	54.8
Causalainer	**68.2**	**55.5**

224 with RGB channels. Every channel is normalized by standard deviation $= (0.2737, 0.2631, 0.2601)$ and mean $= (0.4280, 0.4106, 0.3589)$. PyTorch and NVIDIA TITAN Xp GPU are used for the implementation and to train models for 60 epochs with $1e-6$ learning rate. The Adam optimizer is used [40], with hyper-parameters set as $\epsilon = 1e-8$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

Causal Learning Dataset. When we observe people's writing behaviors, we notice some of them happen very often, such as synonym replacement, accidentally missing some words in a sentence, and so on. Motivated by the above, we randomly pick up one of the behaviors, e.g., accidentally missing some words in a sentence, and write a textual intervention function to simulate it. Similarly, we know that when people make videos in their daily life, some visual disturbances may exist, e.g., salt and pepper noise, image masking, blurring, and so on. We also randomly pick up some of them, e.g., blur and salt and pepper noise, and make a visual intervention function to do the simulation. Based on the visual and textual simulation functions, we can make our causal video summarization dataset with visual and textual interventions. The dataset is made based on the following steps. First, 50% of the *(video, query)* data pairs are randomly selected from the original training, validation, and testing sets. Secondly, for each selected video, 0 or 1 intervention labels are randomly assigned to 30% of the video frames and the corresponding queries. Note that in real-world scenarios, there are various disturbances beyond the previously mentioned visual and textual interventions that could be utilized in the proposed method.

3.2 Evaluation and Analysis

Evaluation protocol. Following existing works [13, 26, 27, 37, 55], we evaluate the proposed method under the same setting. TVSum, QueryVS, and SumMe datasets are randomly divided into five splits, respectively. For each of them, 80% of the dataset is used for training, and the remaining for evaluation. F_1-score [13, 18, 37, 55] is adopted to measure the matching degree of the generated video summaries $\mathbb{S}_i$ and the ground-truth video summaries $\hat{\mathbb{S}}_i$ for video i.

Table 3. Comparison with weakly-supervised state-of-the-art methods. The performance of the proposed approach is better than the existing weakly-supervised method.

Weakly-supervised Method	TVSum
Random summary	54.4
WS-HRL [6]	58.4
Causalainer	**66.9**

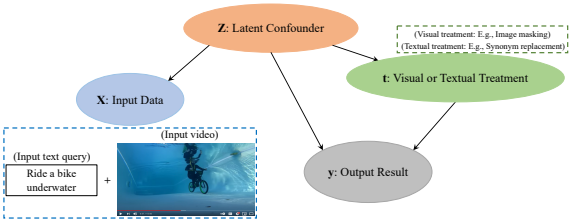

Figure 3. Causal graph in video summarization. **t** is an intervention, e.g., visual or textual perturbation. **y** is an outcome, e.g., an importance score of a video frame or a relevance score between the input text query and video. **Z** is an unobserved confounder, e.g., representativeness, interestingness, or storyline smoothness. **X** is noisy views on the hidden confounder **Z**, say the input text query and video. The causality graph of video summarization leads to more explainable modeling.

State-of-the-art comparisons. The proposed method outperforms existing state-of-the-art (SOTA) models based on different supervision schemes, as shown in Table 1, Table 2, and Table 3. This is because the introduced causal modeling strengthens the causal inference ability of a video summarization model by uncovering the causal relations that guide the process and result.

Effectiveness analysis of the proposed causal modeling. The proposed approach differs from existing methods by introducing causal modeling. Hence, the results in Tables 1, 2, and 3, demonstrate the effectiveness of this approach and serve as an ablation study of causal learning. An auxiliary task/distribution is a key component of the proposed approach, helping the model learn to diagnose input to make correct inferences for the main task, i.e., video summary inference, During training, a binary causation label is provided to teach the model to perform well regardless of intervention. This implies the model has the ability to analyze input and perform well in the main task, making it more robust.

Explainability improvement analysis. The Causalainer method benefits modeling explainability with its associated causal graph of video summarization. Latent factors affecting video summary generation are treated as the causal effect in the proposed causal modeling. A causal graphical model is used to approach the video summarization problem, and the modeling explainability is illustrated in Figure 3.

4. Conclusion

ML-based decision-making systems, like video summarization, suffer from a lack of explainability, resulting in mistrust. To improve modeling explainability, we propose a new Causalainer method that achieves state-of-the-art F_1-score performance in video summarization.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054, 2020. 2

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020. 2

[3] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *arXiv preprint arXiv:2101.06072*, 2021. 3

[4] Sijia Cai, Wangmeng Zuo, Larry S Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 184–200, 2018. 1, 3

[5] Zhihong Cai and Manabu Kuroki. On identifying total effects in the presence of latent variables and selection bias. *arXiv preprint arXiv:1206.3239*, 2012. 2

[6] Yiyan Chen, Li Tao, Xueting Wang, and Toshihiko Yamasaki. Weakly supervised video summarization by hierarchical reinforcement learning. In *Proceedings of the ACM Multimedia Asia*, 2019. 1, 3, 4

[7] Riccardo Di Sipio, Jia-Hong Huang, Samuel Yen-Chi Chen, Stefano Mangini, and Marcel Worring. The dawn of quantum natural language processing. *ICASSP*, 2022. 2

[8] Mohamed Elfeki and Ali Borji. Video summarization via actionness ranking. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 754–763. IEEE, 2019. 4

[9] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian Conference on Computer Vision*, pages 39–54. Springer, 2018. 4

[10] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in neural information processing systems*, pages 2069–2077, 2014. 1

[11] Sander Greenland and DAVID G KLEINBAUM. Correcting for misclassification in two-way tables and matched-pair studies. *International Journal of Epidemiology*, 12(1):93–97, 1983. 2

[12] Sander Greenland and Timothy L Lash. Bias analysis. *International Encyclopedia of Statistical Science*, 2:145–148, 2011. 2

[13] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, pages 505–520. Springer, 2014. 1, 3, 4

[14] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3090–3098, 2015. 1

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[17] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018. 1

[18] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298, 2005. 4

[19] Tao Hu, Pascal Mettes, Jia-Hong Huang, and Cees GM Snoek. Silco: Show a few images, localize the common object. In *ICCV*, pages 5067–5076, 2019. 2

[20] Jia-Hong Huang. Robustness analysis of visual question answering models by basic questions. *King Abdullah University of Science and Technology, Master Thesis*, 2017. 2

[21] Jia-Hong Huang, Modar Alfadly, and Bernard Ghanem. Vqabq: Visual question answering by basic questions. *VQA Challenge Workshop, CVPR*, 2017. 2

[22] Jia-Hong Huang, Modar Alfadly, and Bernard Ghanem. Robustness analysis of visual qa models by basic questions. *VQA Challenge and Visual Dialog Workshop, CVPR*, 2018. 2

[23] Jia-Hong Huang, Modar Alfadly, Bernard Ghanem, and Marcel Worring. Assessing the robustness of visual question answering. *arXiv preprint arXiv:1912.01452*, 2019. 2

[24] Jia-Hong Huang, Modar Alfadly, Bernard Ghanem, and Marcel Worring. Improving visual question answering models through robustness analysis and in-context learning with a chain of basic questions. *arXiv preprint arXiv:2304.03147*, 2023. 2

[25] Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. A novel framework for robustness analysis of visual qa models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 8449–8456, 2019. 2

[26] Jia-Hong Huang, Luka Murn, Marta Mrak, and Marcel Worring. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *ICMR*, pages 580–589, 2021. 1, 4

[27] Jia-Hong Huang and Marcel Worring. Query-controllable video summarization. In *ICMR*, pages 242–250, 2020. 1, 3, 4

[28] Jia-Hong Huang, Ting-Wei Wu, and Marcel Worring. Contextualized keyword representations for multi-modal retinal image captioning. In *ICMR*, pages 645–652, 2021. 2

[29] Jia-Hong Huang, Ting-Wei Wu, C-H Huck Yang, Zenglin Shi, I Lin, Jesper Tegner, Marcel Worring, et al. Non-local attention improves description generation for retinal images. In *WACV*, pages 1606–1615, 2022. 2

[30] Jia-Hong Huang, Ting-Wei Wu, Chao-Han Huck Yang, and Marcel Worring. Deep context-encoding network for retinal image captioning. In *ICIP*, pages 3762–3766. IEEE, 2021. 2

[31] Jia-Hong Huang, Ting-Wei Wu, Chao-Han Huck Yang, and Marcel Worring. Longer version for" deep context-encoding network for retinal image captioning". *arXiv preprint arXiv:2105.14538*, 2021. 2

[32] Jia-Hong Huang, Chao-Han Huck Yang, Pin-Yu Chen, Andrew Brown, and Marcel Worring. Causal video summarizer for video exploration. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2

[33] Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, et al. Deepopht: medical report generation for retinal images via deep models and visual explanation. In *WACV*, pages 2442–2452, 2021. 2

[34] C-H Huck Yang, Fangyu Liu, Jia-Hong Huang, Meng Tian, I-Hung Lin, Yi Chieh Liu, Hiromasa Morikawa, Hao-Hsiang Yang, and Jesper Tegner. Auto-classification of retinal diseases in the limit of sparse data using a two-streams machine learning model. In *ACCV*, pages 323–338. Springer, 2018. 2

[35] Zhong Ji, Fang Jiao, Yanwei Pang, and Ling Shao. Deep attentive and semantic preserving video summarization. *Neurocomputing*, 405:200–207, 2020. 1, 4

[36] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 1, 4

[37] Hao Jiang and Yadong Mu. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16388–16398, 2022. 1, 4

[38] Yudong Jiang, Kaixu Cui, Bo Peng, and Changliang Xu. Comprehensive video understanding: Video summarization with content-based video recommender design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[39] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, 2021. 1

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[42] Jie Lei, Qiao Luan, Xinhui Song, Xiao Liu, Dapeng Tao, and Mingli Song. Action parsing-driven video summarization based on reinforcement learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):2126–2137, 2018. 1

[43] Yujie Li, Atsunori Kanemura, Hideki Asoh, Taiki Miyanishi, and Motoaki Kawanabe. Extracting key frames from first-person videos in the common space of multiple sensors. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3993–3997. IEEE, 2017. 1

[44] Yi-Chieh Liu, Hao-Hsiang Yang, C-H Huck Yang, Jia-Hong Huang, Meng Tian, Hiromasa Morikawa, Yi-Chang James Tsai, and Jesper Tegner. Synthesizing new retinal symptom images by multiple generative models. In *ACCV*, pages 235–250. Springer, 2018. 2

[45] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NIPS*, pages 6446–6456, 2017. 2

[46] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Video summarization using deep semantic features. In *Asian Conference on Computer Vision*, pages 361–377. Springer, 2016. 1

[47] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3657–3666, 2017. 1

[48] Judea Pearl. Bayesianism and causality, or, why i am only a half-bayesian. In *Foundations of bayesianism*, pages 19–36. Springer, 2001. 1

[49] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018. 1, 2

[50] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 2

[51] Melissa Sanabria, Frédéric Precioso, and Thomas Menguy. A deep architecture for multimodal summarization of soccer games. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pages 16–24, 2019. 1

[52] Jan Selén. Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *Journal of the American Statistical Association*, 81(393):75–81, 1986. 2

[53] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017. 2

[54] Xinhui Song, Ke Chen, Jie Lei, Li Sun, Zhiyuan Wang, Lei Xie, and Mingli Song. Category driven deep recurrent neural network for video summarization. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016. 1

[55] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 1, 3, 4

[56] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 582–590, 2017. 1

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[58] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. Video summarization via semantic attended networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 4

[59] Ting-Wei Wu, Jia-Hong Huang, Joseph Lin, and Marcel Worring. Expert-defined keywords improve interpretability of retinal image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868, 2023. 2

[60] Xiang Yan, Syed Zulqarnain Gilani, Mingtao Feng, Liang Zhang, Hanlin Qin, and Ajmal Mian. Self-supervised learning to detect key frames in videos. *Sensors*, 20(23):6941, 2020. 1

[61] C-H Huck Yang, Jia-Hong Huang, Fangyu Liu, Fang-Yi Chiu, Mengya Gao, Weifeng Lyu, Jesper Tegner, et al. A novel hybrid machine learning model for auto-classification of retinal diseases. *Workshop on Computational Biology, ICML*, 2018. 2

[62] Chao-Han Huck Yang, I Hung, Te Danny, Yi Ouyang, and Pin-Yu Chen. Causal inference q-network: Toward resilient reinforcement learning. *arXiv preprint arXiv:2102.09677*, 2021. 2

[63] Chao-Han Huck Yang, I-Te Hung, Yi-Chieh Liu, and Pin-Yu Chen. Treatment learning causal transformer for noisy image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6139–6150, 2023. 2

[64] Yuan Yuan, Haopeng Li, and Qi Wang. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. *IEEE Access*, 7:64676–64685, 2019. 4

[65] Yitian Yuan, Tao Mei, Peng Cui, and Wenwu Zhu. Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):226–237, 2017. 1, 4

[66] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018. 1, 2

[67] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[68] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, pages 766–782. Springer, 2016. 1, 4

[69] Yujia Zhang, Michael Kampffmeyer, Xiaoguang Zhao, and Min Tan. Dtr-gan: Dilated temporal relational adversarial network for video summarization. In *Proceedings of the ACM Turing Celebration Conference-China*, pages 1–6, 2019. 1

[70] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871, 2017. 1, 4

[71] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7405–7414, 2018. 1

[72] Kaiyang Zhou, Tao Xiang, and Andrea Cavallaro. Video summarisation by classification with deep reinforcement learning. *arXiv preprint arXiv:1807.03089*, 2018. 1, 4