MICap: A Unified Model for Identity-aware Movie Descriptions

Haran Raajesh*[1] Naveen Reddy Desanur*[1] Zeeshan Khan[2] Makarand Tapaswi[1]
[1]CVIT, IIIT Hyderabad, India
[2]Inria Paris and Département d'informatique de l'ENS, CNRS, PSL Research University
https://katha-ai.github.io/projects/micap/
* denotes equal contribution

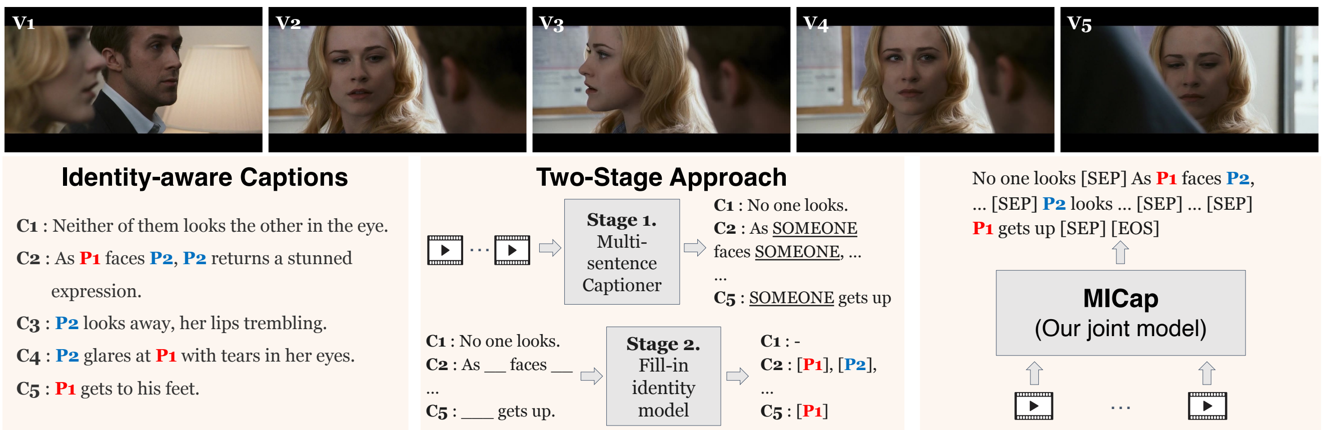

Figure 1. Identity-aware captioning. **Left:** To understand the story in a set of videos, captions refer to characters by a unique local identifier (*e.g.* P1, P2, ...). The *Fill-in-the-blanks* (FITB) task provides these captions with blanks (removing names) and asks a model to fill local person ids. **Middle:** End-to-end captioning for a videoset is achieved in two stages [29]. First, captions are generated with *someone*tags, and then the FITB module is applied to fill-in names. **Right:** We propose a single-stage encoder-decoder id-aware captioning approach that can switch between generating the caption with ids or filling in the ids in a caption, jointly learning from both tasks.

Abstract

Characters are an important aspect of any storyline and identifying and including them in descriptions is necessary for story understanding. While previous work has largely ignored identity and generated captions with someone *(anonymized names), recent work formulates id-aware captioning as a fill-in-the-blanks (FITB) task, where, given a caption with blanks, the goal is to predict person id labels. However, to predict captions with ids, a two-stage approach is required: first predict captions with* someone*, then fill in identities. In this work, we present a new single stage approach that can seamlessly switch between id-aware caption generation or FITB when given a caption with blanks. Our model, Movie-Identity Captioner (MICap), uses a shared auto-regressive decoder that benefits from training with FITB and full-caption generation objectives, while the encoder can benefit from or disregard captions with blanks as input. Another challenge with id-aware captioning is the lack of a metric to capture subtle differences between person ids. To this end, we introduce iSPICE, a caption evaluation metric that focuses on identity tuples created through intermediate scene graphs. We evaluate MICap on Large-Scale Movie Description Challenge (LSMDC), where we show a 4.2% improvement in FITB accuracy, and a 1-2% bump in classic captioning metrics.*

1. Introduction

Building computer vision models that understand the story of a movie is a long-standing challenge. A step towards this is movie description [30, 37, 38]. Given a short clip of 2-5 seconds, models are required to generate a caption that describes the visual scene. Captions in the Large Scale Movie Description Challenge (LSMDC) [38], a combination of [30, 37], are obtained from *audio descriptions* (AD) that are used to convey the (visual) story to a visually impaired audience. The original version of the LSMDC challenge suggests captioning a single clip and anonymizes all character names with *someone*.

While using the *someone* tag to describe a character's activity in a single video is acceptable, the lack of identity continuity across a *videoset* (group of N consecutive videos) hampers understanding. To remedy this, Pini *et al*. [31] extend MVAD [30] as *MVAD names* where character names are predicted by linking to the appropriate face detection/track; and Park *et al*. [29] propose a fill-in-the-blanks (FITB) task to replace *someone* tags with local cluster identities (*e.g*. P1, P2, . . .) in a videoset (Fig. 1 left).

The latter approach [29] provides two advantages: (i) it does not require time-consuming ground-truth annotations linking faces and blanks [31]; and (ii) using local cluster ids helps convey the story[1] without the need for models with world knowledge (CLIP [33], GPT [32], *etc*.) or an IMDb castlist with photographs [14], making the approach applicable to indie films or home-edited videos.

To generate id-aware captions, [29] proposes a two-stage approach shown in Fig. 1 (middle). The first stage [28] ingests a videoset and generates a *captionset* (a set of N captions, one for each video) using the *someone* tags; while the second stage replaces *someone* with appropriate local person id labels. While the two-stage setting unites the two worlds of video description and character identification, it is not ideal as errors in captioning may adversely affect FITB as both methods are modeled independently. In this work, we propose a single-stage approach (Fig. 1 right) that can seamlessly switch between both tasks.

Challenges with Fill-In. For the FITB task, [29] encodes blanks in the ground-truth (GT) captionset using bidirectional context through the BERT encoder. These blanks attend to the face features clustered within a single video, not accounting for other faces coming from the videoset. Using the blank representations, the person ids are predicted in an auto-regressive manner. We note some disadvantages with this approach: (i) Faces are clustered within each video. This means identity information across videos is not directly observed by the model. (ii) When a character is mentioned in the caption, their face need not be present in the clip (*e.g*. Fig. 1 left, C4 and C5 mention P1 whose face is turned and not visible). (iii) BERT-based blank embeddings provided at the encoder are unable to capture face information properly, resulting in a model that largely focuses on text embeddings to solve FITB (*e.g*., in [29], FITB accuracy only improves by 1.5% (64.4 to 65.9) with visual inputs).

Proposed model benefits. We overcome these problems using a new paradigm for id-aware multi-video description through a single-step sequence-to-sequence model. We unify the two tasks of FITB and caption generation, by auto-regressively unrolling the descriptions along with their local character ids, via a Transformer based encoder-decoder model. Our model, dubbed as the *Movie-Identity Captioner* (MICap), enables joint training and independent evaluation for both tasks: (i) given only the videoset, our model generates an id-aware captionset; and (ii) when a captionset with *someone* tags exists, our model fills in local identities.

To overcome text-only shortcuts, we propose auto-regressive decoding of the full caption even for FITB and show that our multimodal model outperforms a text-only model significantly. We teacher force the ground-truth caption containing the blanks (person ids), and predict one token at a time using causal masking. Note, learning happens only at select tokens where person id labels are predicted. This way the model (decoder) learns to sequentially use the GT (teacher forced) caption for the FITB task with unidirectional (causal) attention. During inference, we switch between the two tasks by deciding whether the decoder is teacher forced with a given captionset or not.

Identity-aware evaluation. Existing captioning metrics like CIDEr [50] and BLEU [27] do not account for identity sensitive descriptions. For example *"P1 is walking towards P2"* and *"P2 is walking towards P1"* will result in high n-gram based scores due to common middle words. We propose a new identity-aware caption evaluation metric *iSPICE*. Specifically, we are motivated by SPICE's [1] ability to parse a caption into a scene graph, and match a predicted caption with ground-truth based on similarity across generated tuples. To compute iSPICE, we intervene in this process and remove tuples not associated with a person label before computing the F1 scores.

Contributions. In summary, (i) we propose a new paradigm for identity-aware multi-sentence movie description using a single-stage approach that unifies FITB with full caption generation. (ii) We formulate this task as an auto-regressive sequence-to-sequence generation that is able to describe the video and use local person id labels across a videoset (multiple videos). We show that joint training improves knowledge sharing and boosts performance. (iii) We enable seamless task switching allowing independent evaluation of (a) caption generation with identities, and (b) filling in identity labels given a caption. (iv) We propose a new identity-aware captioning metric, iSPICE, that extends SPICE, and show its sensitivity to identities while evaluating captions. (v) Finally, MICap improves over the state-of-the-art for FITB by 4.2% and identity-aware captioning by 1.4% CIDEr and 1.8% METEOR.

2. Related Work

We address related work from three areas: (i) video captioning at large, (ii) identity-aware captioning, and (iii) metrics used for evaluating captions.

Video captioning has gained a lot of attention since the ad-

[1] Note, cluster ids can be easily mapped to gender- and culture-appropriate names instead of using P1, P2, . . . for storytelling.

vent of deep learning. The typical task is to generate a single sentence description for a trimmed video, and is formulated as a sequence-to-sequence problem [12, 22, 23, 42, 51, 52, 58]. A more challenging setup is multi-sentence generation, typically applied to longer videos and requires long-term temporal consistency [28, 36, 45, 57]. Video situation recognition, VidSitu [17, 39] presents a structured alternative where multiple captions are generated per event based on the semantic role labeling framework.

Different from multi-sentence captioning, *dense video captioning*, requires temporally localizing and generating captions for every event in an untrimmed video [18, 55, 56, 62]. While most approaches for dense video captioning use a 2-stage approach, *i.e*. temporal localization with event proposals then event captioning [18, 53, 54], recent methods, jointly model the two tasks for better temporal consistency [5, 7, 8, 20, 25, 35, 43, 44, 53, 55, 62]. The state-of-the-art, PDVC [55], learns DETR-style event queries and performs localization and captioning over each query using 2 separate heads. Recently, Vid2Seq [56] proposed to further unify the two tasks by using a single sequence-to-sequence model and generating both the localization and captions with a single auto-regressive Transformer decoder. Similar to above ideas, we unify two seemingly different tasks of character identification and description by formulating them as an auto-regressive sequence generation task.

Id-aware captioning datasets. None of the above works focus on person identity while generating captions. VidSitu [39], perhaps the closest, contains references to people by descriptions such as *man in a black jacket*. This is an issue when the domain is movie description [30, 38], where identities are anonymized to *someone*which hinders building practical applications like *Audio Descriptions* [13] for visually impaired users. While [31] links character names in descriptions with face tracks, they require significant annotation effort that is not scalable. A more recent Movie Audio Description dataset, MAD [46], is a popular source for movie descriptions. But it uses real names that require models with world knowledge. Different from above, Park *et al*. [29] propose identity-aware captioning as a fill-in-the-blanks task where they assign local person ids (cluster ids) to characters appearing in 5 consecutive video clips. We adopt this setting for our work.

Id-aware captioning methods. Identity-aware captioning is a challenging task that has recently started to attract attention. Among the first works, [29] proposes a 2-stage pipeline of first captioning with identities anonymized as *someone* using a multi-sentence captioning model [28], followed by learning an identity prediction FITB model that fills in the *someone* with local person identities. However, as discussed in the introduction (Challenges with Fill-In), the specific 2-stage approach suffers from several disadvantages. Different from [29], we propose a single stage sequence-to-sequence model, that outperforms the 2-stage approach. In this area, another work [60] requires ground-truth mapping between person identities (blanks) in the description to face tracks in the videos. However, this approach is not scalable. Very recently, AutoAD-II [14] proposed to generate movie descriptions with proper names, on the MAD [46] dataset. While innovative, this approach requires additional IMDb castlist information with photographs. While modeling proper names directly is useful, tagging names to unique person ids in a local videoset is possible and is the motivation for works on person clustering [3, 48] as opposed to person identification [26, 47].

Caption evaluation metrics are typically based on n-gram matching, with few differences. CIDEr [50], BLEU [27], and METEOR [11] all evaluate n-gram similarities between a single or multiple candidate references and the generated caption. Recently, Large Language Models (LLMs) are used for reference-based (*e.g.* BERTScore [61], CLAIR [6]) or or Large Vision-Language Models (VLMs) for reference-free caption evaluation (*e.g.* CLIP Score [15]). However, model-based metrics may be difficult to interpret, and also require the model to be sensitive to identities. Different from both directions, SPICE [1] evaluates captions by first transforming them into a scene graph and analyzing presence of shared tuples between the predicted and ground-truth (reference) captions. However, none of the metrics reliably evaluate identity-aware captions, as a robust metric should be sensitive to identity manipulations (swap/add/remove). We propose a new metric iSPICE that focuses primarily on person-identity specific semantics.

3. Method

We present a single-stage sequence-to-sequence approach for identity-aware fill-in-the-blanks (FITB). Later, we will show that this architecture can be easily re-purposed for generating video descriptions.

Notation. Before we start, we define some notation. For the rest of this section, we will operate with a videoset $\mathcal{N}$ consisting of N video clips V_i and corresponding captionset $\mathcal{C} = \{C_i\}_{i=1}^{N}$, where C_i describes video V_i. As both sets come from consecutive videos, it is very likely that same characters appear across them. As an example, consider the videoset frames and captionset shown in Fig. 1.

3.1. Auto-regressive FITB

In FITB, we replace each person-id (P1, P2, . . .) with a blank. We denote $\hat{\mathcal{C}}$ as the captionset with $\mathcal{B}$ blanks. Formally, we define the captionset as a sequence of L words $[w_j]_{j=1}^{L}$, some of which have been converted to blanks $\{b_k\}_{k=1}^{|\mathcal{B}|}$. The goal of our model is to fill each blank with the correct person-id label from the set $\mathcal{P} = \{P_l\}_{l=1}^{|\mathcal{P}|}$. Note, the person-id labels are reusable across videosets, *i.e*. a charac-

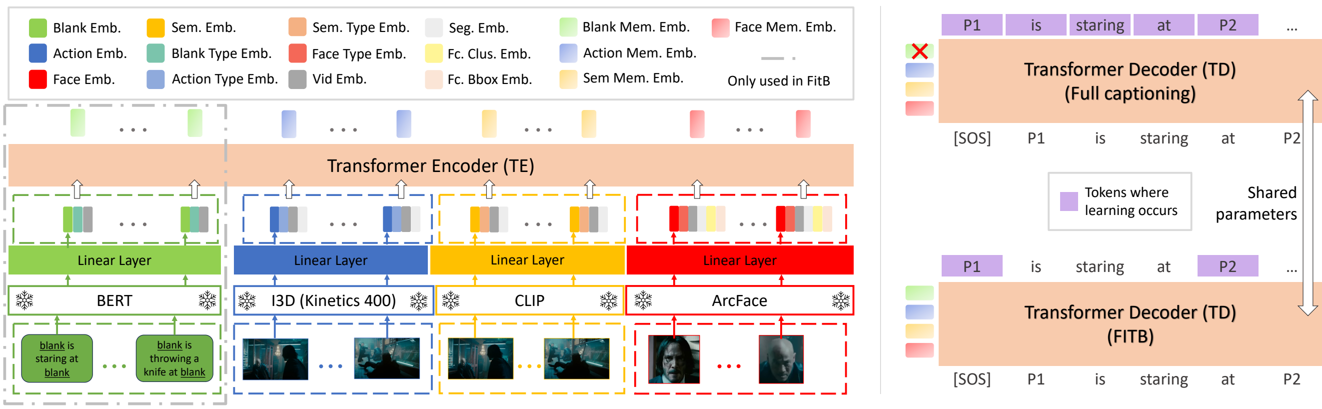

Figure 2. Identity-aware captioning. **Left**: illustrates the Transformer Encoder used to capture multimodal inputs such as text (blanks), action, semantic, and face. These tokens are used as memory for the Transformer Decoders. **Right**: the same Transformer Decoder can be used for both tasks of full caption generation and fill-in-the-blanks (FITB). The model is trained end-to-end with losses applied to tokens indicated in purple. Text tokens are not presented to the decoder for full caption generation. Joint training improves knowledge sharing resulting in performance improvements.

ter only needs to be referred consistently by the same identity within a videoset.

We present Movie-Identity Captioner (MICap), an auto-regressive Transformer encoder-decoder model for filling person blanks. MICap consists of two parts: (i) Feature extractors and a Transformer encoder to build the captioning memory (Fig. 2 left); and (ii) A Transformer decoder that switches between FITB or full captionset generation (Fig. 2 right). For clarity, we will highlight differences to prior work [29] throughout this section.

3.1.1 Creating the Captioning Memory

Visual feature extraction. We extract 3 features from the videoset to capture semantic, action, and face information.

Semantic embeddings are captured using CLIP [33]. From each video V_i, we sub-sample frames f_{it} at 5 fps and encode them with the CLIP image encoder. For efficient batching, we truncate or pad to $T{=}50$ frames per video, and stack them to create semantic features $\mathbf{F}^{\text{s}} \in \mathbb{R}^{NT \times d^{\text{s}}}$.

Action embeddings are captured using I3D [4]. Similar to [29], each video is divided into $S{=}5$ segments, and features within each segment are mean pooled. We stack features across the videoset to obtain $\mathbf{F}^{\text{a}} \in \mathbb{R}^{NS \times d^{\text{a}}}$.

Faces are detected using Retina Face [10] and represented using Arcface [9]. Across the videoset, we collect a maximum of $F{=}300$ face detections. With each face detection, we associate the video index i (for V_i) from which it is derived and a normalized spatial bounding box location. We stack features to obtain $\mathbf{F}^{\text{f}} \in \mathbb{R}^{F \times d^{\text{f}}}$.

We bring all these features to a common d dimensional space using separate linear projection layers for each modality: $\mathbf{W}^{\text{mod}} \in \mathbb{R}^{d \times d^{\text{mod}}}$, where mod takes on values: s for semantic, a for action, and f for face.

Captionset feature extraction. Similar to [29], we also extract blank embeddings by feeding the captionset to BERT (fine-tuned for gender prediction as in [29]) and using the contextualized tokens:

$$[\hat{\text{CLS}}, \hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{b}}_k, \ldots] = \text{BERT}([\text{CLS}, w_1, \ldots, b_k, \ldots])\,. \quad (1)$$

The blank embedding is a concatenation of contextualized tokens: $\mathbf{b}_k = [\hat{\text{CLS}}, \hat{\mathbf{b}}_k]$. We stack these to create a matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{B}| \times 2 \cdot d^{\text{bert}}}$ and transform them to the same space through a linear projection $\mathbf{W}^{\text{bert}} \in \mathbb{R}^{d \times 2 \cdot d^{\text{bert}}}$.

Face clustering. Instead of creating face clusters within each video and using blank embeddings to attend to them (as done in [29]) we adopt a soft approach for incorporating cluster information in MICap. First, we perform clustering using DBSCAN across *all* F detections in the *videoset*, resulting in $\mathcal{G}$, a set of face groups. This allows our model to associate faces across videos as the same or different person. Next, we prevent propagating errors caused by clustering and mean pooling representations by adding a cluster-id based learnable embedding $\mathbf{E}^{\text{fcl}}$ to the face representations.

Additional embeddings are added to various features to orient the model: (i) $\mathbf{E}^{\text{typ}} \in \mathbb{R}^{d \times 4}$ disambiguates between the 4 types of features. (ii) $\mathbf{E}^{\text{vid}} \in \mathbb{R}^{d \times N}$ consists of N embeddings to inform the model of the source video index for any visual or blank token. (iii) $\mathbf{E}^{\text{seg}} \in \mathbb{R}^{d \times S}$, together with $\mathbf{E}^{\text{vid}}$, allows to localize any feature to the correct video and segment. (iv) $\mathbf{E}^{\text{fcl}} \in \mathbb{R}^{d \times |\mathcal{G}|}$ is the face cluster index embedding described above, and (v) $\mathbf{E}^{\text{bbox}} \in \mathbb{R}^{d \times 4}$ transforms normalized face detection bounding box coordinates to provide the model spatial information.

We create input tokens as follows (with appropriate indexing hidden for brevity):

$$\hat{\mathbf{B}} = \mathbf{W}^{\text{bert}}\mathbf{B} + \mathbf{E}_0^{\text{typ}} + \mathbf{E}^{\text{vid}}\,, \tag{2}$$

$$\hat{\mathbf{F}}^{\text{s}} = \mathbf{W}^{\text{s}}\mathbf{F}^{\text{s}} + \mathbf{E}_1^{\text{typ}} + \mathbf{E}^{\text{vid}} + \mathbf{E}^{\text{seg}}\,, \tag{3}$$

$$\hat{\mathbf{F}}^{\text{a}} = \mathbf{W}^{\text{a}}\mathbf{F}^{\text{a}} + \mathbf{E}_2^{\text{typ}} + \mathbf{E}^{\text{vid}} + \mathbf{E}^{\text{seg}}\,, \tag{4}$$

$$\hat{\mathbf{F}}^{\text{f}} = \mathbf{W}^{\text{f}}\mathbf{F}^{\text{f}} + \mathbf{E}_3^{\text{typ}} + \mathbf{E}^{\text{vid}} + \mathbf{E}^{\text{seg}} + \mathbf{E}^{\text{fcl}} + \mathbf{E}^{\text{bbox}}. \tag{5}$$

A Transformer encoder (TE) [49] of L_E layers is used to combine and refine individual representations mentioned above. Thus, the final memory bank is:

$$\mathbf{M} = [\tilde{\mathbf{B}}, \tilde{\mathbf{F}}^s, \tilde{\mathbf{F}}^a, \tilde{\mathbf{F}}^f] = \text{TE}([\hat{\mathbf{B}}, \hat{\mathbf{F}}^s, \hat{\mathbf{F}}^a, \hat{\mathbf{F}}^f])\,. \tag{6}$$

3.1.2 Auto-regressive Identity Prediction

We now present the process of filling blanks. Similar to the encoder, we use a couple embeddings for the decoder. (i) $\mathbf{E}^{\text{vid}}$ (shared with encoder) informs the decoder of the video index that is being captioned; and (ii) $\mathbf{E}^{\text{pos}}$ encodes learnable position embeddings similar to the original Transformer [49]. We use the memory embeddings extracted from the video as key-value pairs and blanks in the Transformer decoder (TD) as queries. Given a captionset $\hat{\mathcal{C}}$, we generate the next word as

$$\mathbf{h}_{j+1} = \text{TD}([w_1, \ldots, w_j]; \mathbf{M})\,, \tag{7}$$

$$w_{j+1} = \arg\max_{\mathcal{V}} \mathbf{W}^{\mathcal{V}}\mathbf{h}_{j+1}\,. \tag{8}$$

$\mathbf{h}_{j+1}$ represents the output of TD at the $j+1^{\text{th}}$ timestep and is obtained through a series of L_D decoder layers that compute self-attention to previous words, and cross-attention to the memory. $\mathbf{W}^{\mathcal{V}}$ is a linear classifier in $\mathbb{R}^{\mathcal{V}\times d}$, where $\mathcal{V}$ is the word vocabulary.

For the FITB task, the captionset already contains the correct caption words. Thus, the output prediction is relevant only when w_{j+1} is a blank b_k. In such a case, we can use a smaller output classifier $\mathbf{W}^{\mathcal{P}}$ that picks one among $\mathcal{P}$ person-id labels. We rewrite the above equations as:

$$\mathbf{h}_{j+1} = \text{TD}([w_1, \ldots, w_j]; \mathbf{M})\,, \tag{9}$$

$$w_{j+1} {=} \hat{y}_k = \arg\max_{\mathcal{P}} \mathbf{W}^{\mathcal{P}}\mathbf{h}_{j+1}\,, \tag{10}$$

where $\hat{y}_k \in \mathcal{P}$ is the predicted person-id label for blank b_k.

Training and inference. We train MICap by applying a cross-entropy loss at every blank:

$$\mathcal{L}_{\text{FITB}} = -\sum_{k=1}^{|\mathcal{B}|} y_k \log \text{softmax}_{\mathcal{P}} \left(\mathbf{W}^{\mathcal{P}}\mathbf{h}_{j+1}\right)\,, \tag{11}$$

where y_k is the correct label for blank b_k. The key difference to [29] is that our decoder observes each word of the captionset in an auto-regressive manner.

During inference, we simply follow Eq. (10) to compute person-id label predictions for blanks in a captionset.

3.2. Joint Fill-in and Captioning

We first present how MICap can be adapted for generating the entire captionset. Then, we will present the opportunity of joint training.

From FITB to generating the captionset. In this scenario, the model is shown the videoset $\mathcal{N}$ and expected to generate an id-aware captionset $\mathcal{C}$. We make two small changes:

(i) The memory bank is restricted to visual features, $\mathbf{M} = [\tilde{\mathbf{F}}^s, \tilde{\mathbf{F}}^a, \tilde{\mathbf{F}}^f]$. In fact, we cannot compute blank embeddings $\tilde{\mathbf{B}}$ as the captionset needs to be predicted.

(ii) When decoding the next word of the captionset, we use an augmented vocabulary consisting of normal language tokens (from $\mathcal{V}$) and person-id labels (from $\mathcal{P}$). We predict the next word as shown below:

$$\mathcal{V}^* = \mathcal{V} + \mathcal{P}\,, \tag{12}$$

$$\mathbf{h}_{j+1} = \text{TD}([w_1, \ldots, w_j]; \mathbf{M})\,, \tag{13}$$

$$\hat{w}_{j+1} = \arg\max_{\mathcal{V}^*} \mathbf{W}^{\mathcal{V}^*}\mathbf{h}_{j+1}\,, \tag{14}$$

and train our model to minimize

$$\mathcal{L}_{cap} = -\sum_{j=1}^{L} w_{j+1} \log \text{softmax}_{\mathcal{V}^*} \left(\mathbf{W}^{\mathcal{V}^*}\mathbf{h}_{j+1}\right)\,. \tag{15}$$

We can use Eq. (14) during inference to predict the entire captionset until the end-of-sentence token is triggered.

Joint training. Can we train the same instance of MICap to generate the captionset and fill-in-the-blanks with identity information? Yes, we suggest an efficient way to do so.

Given a batch of data consisting of multiple paired videosets and captionsets $(\mathcal{N}, \mathcal{C})$, we forward it through the model twice. In the first forward pass, we replace the person-id labels with blanks, *i.e.* create $\hat{\mathcal{C}}$, and compute losses and gradients to predict the blank's labels (see Eq. (11)). In the second forward pass conducted on the same batch, we assume that $\mathcal{C}$ is not available as input and use the augmented vocabulary $\mathcal{V}^*$ to compute loss and gradients for each word as in Eq. (15). We can either accumulate gradients and optimize parameters at the end of both forward passes or optimize parameters after each pass.

Note, the classifier parameters $\mathbf{W}^{\mathcal{P}}$ are subsumed under $\mathbf{W}^{\mathcal{V}^*}$. We find that sharing the classifier $\mathbf{W}^{\mathcal{V}^*}$ for both forward passes works best.

Thus, we unite seemingly disparate tasks of filling in person-id labels in blanks and generating the full captionset in a single model with a single set of parameters.

4. Identity-aware SPICE

Inspired by a metric used in image captioning evaluation called Semantic Propositional Image Caption Evaluation (SPICE) [1], we propose a new metric – identity-aware

SPICE (iSPICE for short) – to evaluate the quality of video descriptions, especially pertaining to identity labels.

Why SPICE? The classic captioning metrics borrowed from language translation such as BLEU [27], ROUGE [21], METEOR [11], and CIDEr [50] rely primarily on n-gram overlap. However, as indicated in [1], "n-gram overlap is neither necessary nor sufficient for two sentences to convey the same meaning". SPICE is shown to have a high correlation with human judgement (0.88) as compared to METEOR (0.53) or CIDEr (0.43) on the MS-COCO image captioning dataset [1].

How is SPICE calculated? SPICE estimates quality of a caption in two stages. First, the reference and predicted caption are converted to *scene graphs* [16, 41] that explicitly encode objects, attributes, and relationships. This abstraction provides a list of tuples $\mathcal{T}_r$ and $\mathcal{T}_p$ for the reference and predicted captions. SPICE is the F1-score that measures logical conjunction (overlap):

$$\mathrm{SPICE} = \mathrm{F}_1(\mathcal{T}_r, \mathcal{T}_p)\,. \tag{16}$$

iSPICE is a simple modification of SPICE. We intervene at the list of tuples and filter out tuples that do not have at least one character identity. We define

$$\mathrm{iSPICE} = \mathrm{F}_1(\mathcal{T}_r^{p2+}, \mathcal{T}_p^{p2+}) \cdot \mathrm{F}_1(\mathcal{T}_r^{p1}, \mathcal{T}_p^{p1})\,, \tag{17}$$

where $\mathcal{T}_r^{p2+}$ denotes the list of tuples with a person-id label having 2 or more elements and $\mathcal{T}_r^{p1}$ is a set of person-id labels in the reference captionset. The first term scores whether the correct person-id label is used together with a verb or attribute, while the second term checks that the total number of person-id labels match. A couple examples of the matching process are presented in the supplement.

Validation. We validate iSPICE by an experiment that measures sensitivity to changes in identity. Given a reference captionset, we compare it against itself to obtain a base score s. Next, we modify the reference captionset by swapping, adding new, or removing existing id labels.

1. Swapping: Here, id tokens are replaced with another id present in the captionset. The number of these tokens is selected at random for each captionset. We first identify *eligible id* tokens whose ids are present more than once in the captionset. This is done to prevent the case where standalone ids are selected and replaced with each other that does not change the meaning. For example, the caption *P1 carries P2* is equivalent to *P2 carries P1* if P1 and P2 are not re-used elsewhere in the captionset. When the id occurs multiple times, *e.g. P1 carries P2. P2 is unconscious*, the replacement *P2 carries P1. P2 is unconscious* changes the meaning of the story. Once these eligible tokens are identified, a random subset is replaced with another id present in the captionset to generate the modified caption.

Experiments	iS	S	B4	C	M	R	BSc
Swapping	**0.55**	0.85	0.87	0.86	0.61	0.95	0.99
Addition	**0.51**	0.86	0.89	0.88	0.6	0.95	0.99
Removal	**0.46**	0.84	0.87	0.86	0.6	0.95	0.99

Table 1. Sensitivity of metrics to id manipulation in the original caption. iSPICE shows highest reduction in performance when replacing, adding, or removing ids, indicating that it is a good metric for id-aware captioning iS=iSPICE, S=SPICE, B4=BLEU4, C=CIDEr, M=METEOR, R=ROUGE, BSc=BERTScore.

2. Addition: Here, we select an id token at random and change it to an id token that is not present in the current captionset, adding new identities. Again, we do not replace tokens whose id appears only once.

3. Removal: Here, we replace a single occurrence id token (chosen at random) with an id token that exists in the captionset, thereby removing the identity.

Id normalization. Prior to scoring, a normalization operation is performed on the captionset. The first unique id label is set to P1, the second to P2 an so on. This ensures that the captionsets *P2 carries P1* or *P4 carries P3*, are treated as the same captionset *P1 carries P2*.

Results. We compute a new score $\hat{s}$ for each edited captionset by comparing it against the reference. We report the drop in performance $\hat{s}/s$ as the sensitivity of a metric to changing identities. We create 3 manipulated samples for each type and report averaged scores over all 1443 captionsets from the validation set in Tab. 1. We observe that iSPICE obtains the smallest score, indicating the highest sensitivity to manipulating identities, a desirable property.

5. Experiments

We present experiments on the LSMDC [38] dataset in the identity-aware multi-video captioning setup [29]. We describe the experimental setup first, followed by implementation details and metrics. The evaluation is presented for (i) Fill-in-the-blanks and (ii) Identity-aware captioning.

5.1. Setup

Dataset. LSMDC consists of 128,118 short video clips extracted from 202 movies. Each video has a caption, either from the movie script or from transcribed DVS (descriptive video services) for the visually impaired. The median video duration is 3 s, average is 4.2 s, and std dev is 3.1 s. The dataset is split into 101,079 clips for training, 7,408 for validation, 10,053 for public test, and 9,578 for blind test. We report and compare results on the validation set as the test set labels are not released and the evaluation server is down.

In the Fill-in challenges, the movie descriptions are evaluated on sets of 5 clips taken at a time. Characters are identified across the clips to provide meaningful narratives. The training videosets use overlapping clips (*e.g.* 1-5, 2-6) for data augmentation but the val and test videosets are non-overlapping. We train on 98,527 videosets and report results on 1,443 val videosets. All three tasks of the LSMDC challenge [38] are evaluated on the same sets of 5 clips. We focus on task 2: filling in local person ids; and task 3: description generation with local character IDs.

Implementation details. Videosets have N=5 clips, we set the captionset length to 120 tokens. The hidden dimension for encoder and decoder in MICap is d=512, and we use L_E=2 and L_D=3 layers. We train our model with a learning rate of 5×10^{-5} for 30 epochs. The vocabulary sizes are $|\mathcal{P}|$=11 and $|\mathcal{V}|$=30522. We train on one RTX 2080 GPU with a batch size of 16 videosets/captionsets.

Fill-in metrics. For the Fill-in task we evaluate results using all pairs of blanks in the captionset as proposed by [29]. Pairs that require both ids to be same are called are evaluated with same accuracy ("Same-acc"). Different id pairs are evaluated using "Diff-acc". "Inst-acc" is the combined accuracy while "Class-acc" computes the harmonic mean.

Captioning metrics. We use METEOR [11], CIDEr [50], SPICE [1] and our newly proposed metric iSPICE to evaluate the quality of our generated captions.

5.2. Evaluating on the Fill-in Task

MICap makes better use of visual features. In Tab. 2, our text-only model (row 2) is comparable to [29]'s text-only (R0). While [29] improves by 1.5% (R1), MICap achieves a significant 4.7% improvement (R6).

Ablations on visual features. [29] computes face clusters within a video and provides mean pooled features of faces in a cluster. R3 of Tab. 2 uses these features in MICap (with embeddings from Eq. (5)). The only decoder model (only-dec) achieves a 0.6% improvement, while the encoder-decoder model (enc-dec) shows 1.4% improvement over R1. Next, in R4, we swap out face cluster features to individual face detections, while still using FaceNet for a fair comparison; but using embeddings as shown in Eq. (5). This improves the only-dec model by a further 0.9%, but enc-dec shows negligible change. We incorporate CLIP features as additional tokens in the memory, resulting in a 0.35% increase in enc-dec (R5). Finally, in R6, swapping FaceNet [40] to Arcface [9] results in a relatively large improvement of 1.6% (only-dec) and 1.4% (enc-dec).

SotA comparison. Tab. 3 reports results on all 4 FITB metrics. As we do not have access to the test set labels and the evaluation server is inactive, we use FillIn's results as a proxy for comparison. First, in the top half, we see that FillIn [29] outperforms other works. In the bottom half, on the validation set, we compare our approach against FillIn showing a significant improvement of 4% on instance accuracy and 3.2% on class accuracy. As we teacher force captions through the decoder, our only decoder model also outperforms [29] by 3% on class accuracy.

#	Method	Only Dec	Enc-Dec
0	FillIn text-only [29]	-	64.4
1	FillIn multimodal [29]	-	65.9
2	MICap text-only	-	64.45
3	MICap w face clusters of [29]	66.56	67.29
4	MICap w raw face detections	67.48	67.35
5	MICap 4 + w CLIP features	67.38	67.70
6	MICap 5 + w Arcface features	**68.94**	**69.14**

Table 2. Ablation study showing the impact of various inputs on the decoder only and encoder + decoder model. We report *class accuracy* as a single metric for comparison.

Method	Same	Different	Instance	Class
Test set				
Yu *et al.* [59]	26.4	87.3	65.9	40.6
Brown *et al.* [2]	33.6	81.0	64.8	47.5
FillIn text-only [29]	56.0	71.2	64.8	62.7
FillIn [29]	60.6	70.0	69.6	64.9
Validation set				
FillIn [29]	63.5	68.4	69.0	65.9
Ours (only-dec)	65.1	**73.3**	73.0	68.94
Ours (enc-dec)	**65.7**	72.9	**73.0**	**69.14**

Table 3. Comparison to SotA on fill-in-the-blanks (FITB, task 2) of the LSMDC challenge.

	Captioning metrics				FITB
Method	C	M	S	iS	Class Acc.
FITB only	-	-	-	-	69.14
Full caption only	8.01	12.29	13.11	0.777	-
Joint training	**9.09**	**12.47**	**13.30**	**0.788**	**70.01**

Table 4. Ablation showing joint training is better than performing FITB or full captioning separately. Captioning metrics are C=CIDEr, M=METEOR, iS=iSPICE, S=SPICE.

5.3. Evaluating Joint Fill-in and Captioning

We evaluate MICap trained jointly for FITB and id-aware caption generation. Tab. 4 shows that joint training on fill-in and captioning improves the performance on both the tasks. Class accuracy on FITB improves by 0.9% and captioning metric CIDEr by 1%. We also see a small 0.01% improvement in iSPICE, which we think is important considering the difficulty of the metric. This suggests that both the tasks are complementary and can help each other in learning a better representation. MICap can seamlessly switch between FITB (id prediction) and full caption generation.

SotA comparison for captioning. We compare against the two-stage baseline [29], while MICap predicts the captions and identities in a single stage. Tab. 5 shows that we improve over [29] across all metrics.

MICap's captions are better. We disentangle identity prediction from caption generation by replacing all person id

Figure 3. We show a qualitative example of our joint training approach. The dataset is highly challenging, with shot changes and dark scenes that are typical in movies. Yet our model is able to perform reasonably well in this example. While the predicted captions (Pred) are different from the ground-truth (GT), they capture the overall meaning. MICap predicts diverse ids correctly in this case and does not overfit to only predicting P1, or P1 and P2. In fact, in the last clip, as P3 turns (indicated in GT), we see P4 sitting at the table (indicated in Pred), which is a correct caption! The last clip also highlights challenges of evaluating captions correctly.

	Captions	Method	C	M	S	iS
1		Same id	7.03	9.41	9.01	0.591
2	Fill-in [29]	All diff ids	7	9.11	12.98	0.202
3		FillIn	7.77	10.68	-	-
4		Same id	8.44	10.9	9.26	0.687
5	MICap	All diff ids	8.74	11.01	13.09	0.264
6		MICap (Joint)	**9.09**	**12.47**	**13.30**	**0.788**

Table 5. We evaluate performance of id-aware captioning against [29], showing improvements across all metrics. Captioning metrics are C=CIDEr, M=METEOR, iS=iSPICE, S=SPICE.

Method	Captioning metrics				FITB
	C	M	S	iS	Class Acc.
MICap	**9.09**	**12.47**	**13.30**	**0.788**	**70.01**
T5 only CLIP	4.9	8.5	7.1	0.755	-
T5 all features	4.5	7.9	6.8	0.723	-
GPT2 only CLIP	3.6	8.7	10.7	0.640	-
GPT2 all features	4.4	8.9	9.2	0.595	-

Table 6. Experiments showing MICap outperforms foundational models T5-Base [34] and GPT2 [32] adapted/fine-tuned for id-aware captioning on the same LSMDC dataset.

labels by the same id or all different ids. This allows us to evaluate captioning performance, independent of identity prediction. We are pleased that our simple encoder-decoder approach outperforms a complex adversarial multi-sentence captioning approach [28] used in stage 1 of [29]. Tab. 5 R1 *vs*. R4, CIDEr goes up from 7.03 to 8.44, and METEOR 9.41 to 10.9. Similar improvements hold for R2 *vs*. R5.

Comparison to VLMs. Tab. 6 shows that MICap outperforms adaptations of T5 (an encoder-decoder framework) and GPT-2 (QFormer prefix tokens like CLIPCap [24] or BLIP2 [19]), fine-tuned for the id-aware captioning task. We suspect that integrating many diverse visual tokens is not trivial for VLMs, resulting in comparable performance when using "only CLIP" or "all features".

Id-aware metric. iSPICE is a challenging metric as it multiplies two F1 scores that penalize when the number of identities are mismatched or tuples incorrect. Tab. 5 shows that iSPICE changes dramatically when using the same id or all different ids. We hope that this metric will inspire future works in this direction of identity-aware captioning.

Attention patterns of MICap's decoder reveal interesting insights. For the task of full captioning, we see that tokens that produce id labels cross-attend more to the face tokens (from memory) while normal word tokens cross-attend to CLIP features. We also analyze the attention patterns in FITB and observe that the model attends to the same clusters when predicting the same labels and also attends to face detections across the videoset (not restricted to faces in a single video). Please refer to the supplement for details.

A qualitative example is shown in Fig. 3. We observe that MICap does a decent job at generating captions (although it is unable to use a rich vocabulary - *smiles* instead of *beams cheerily*). The challenges of caption evaluation are also clear in the last clip. Several more examples for both tasks are shown in the supplement.

6. Conclusion

We proposed a new paradigm for identity-aware movie caption generation. As opposed to the two-stage approach of first captioning with anonymized names and then filling in the identities, we proposed a single-stage method that combines the two tasks via an encoder-decoder sequence-to-sequence generation framework, that can seamlessly switch between (i) full caption generation with identities, or (ii) predict the identities given a caption with anonymized names. We showed that a single auto-regressive model benefits both tasks and shows positive transfer, leading to state-of-the-art performance on the LSMDC challenge. We also proposed an identity-aware captioning metric, iSPICE, that is sensitive to subtle perturbations in identity and robustly evaluates captions.

Acknowledgments. The project was supported by funding from SERB SRG/2023/002544. We thank the Bank of Baroda for partial travel support. We thank Amit Pandey for assisting in early discussions. Makarand Tapaswi thanks support from Google India Faculty Award and Naveen Reddy Desanur from Sensara.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 3, 5, 6, 7

[2] Andrew Brown, Samuel Albanie, Yang Liu, Arsha Nagrani, and Andrew Zisserman. LSMDC v2 Challenge presentation. *3rd Workshop on Closing the Loop Between Vision and Language*, 2019. 7

[3] Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. Face, Body, Voice: Video Person-Clustering with Multiple Modalities. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 3

[4] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[5] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iPerceive: Applying Common-Sense Reasoning to Multi-Modal Dense Video Captioning and Video Question Answering. *Winter Conference on Applications of Computer Vision (WACV)*, pages 1–13, 2021. 3

[6] David M Chan, Suzanne Petryk, Joseph E Gonzalez, and Trevor Darrell. CLAIR: Evaluating Image Captions with Large Language Models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 3

[7] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8425–8435, 2021. 3

[8] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 234–243, 2021. 3

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 4, 7

[10] Deng, Jiankang and Guo, Jia and Ververas, Evangelos and Kotsia, Irene and Zafeiriou, Stefanos. Retinaface: Single-shot multi-level face localisation in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[11] Michael Denkowski and Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *European Chapter of the Association for Computational Linguistics (EACL)*, 2014. 3, 6, 7

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015. 3

[13] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD: Movie Description in Context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18930–18940, 2023. 3

[14] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. AutoAD II: The Sequel-Who, When, and What in Movie Audio Description. In *International Conference on Computer Vision (ICCV)*, pages 13645–13655, 2023. 2, 3

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 3

[16] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image Retrieval using Scene Graphs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015. 6

[17] Zeeshan Khan, CV Jawahar, and Makarand Tapaswi. Grounded Video Situation Recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:8199–8210, 2022. 3

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 3

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, 2023. 8

[20] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7492–7500, 2018. 3

[21] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 6

[22] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17949–17958, 2022. 3

[23] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3

[24] Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint 2111.09734*, 2021. 8

[25] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6588–6597, 2019. 3

[26] Arsha Nagrani and Andrew Zisserman. From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script. In *British Machine Vision Conference (BMVC)*, 2017. 3

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Association of Computational Linguistics (ACL)*, pages 311–318, 2002. 2, 3, 6

[28] Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Adversarial inference for multi-sentence video description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6598–6608, 2019. 2, 3, 8

[29] Jae Sung Park, Trevor Darrell, and Anna Rohrbach. Identity-aware multi-sentence video description. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[30] Stefano Pini, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Towards video captioning with naming: a novel dataset and a multi-modal approach. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19*, pages 384–395. Springer, 2017. 1, 2, 3

[31] Stefano Pini, Marcella Cornia, Federico Bolelli, Lorenzo Baraldi, and Rita Cucchiara. M-VAD names: a dataset for video captioning with naming. *Multimedia Tools and Applications*, 78:14007–14027, 2019. 2, 3

[32] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019. 2, 8

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2, 4

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research (JMLR)*, 21:1–67, 2020. 8

[35] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *International Conference on Computer Vision (ICCV)*, pages 8908–8917, 2019. 3

[36] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition (GCPR)*, pages 184–195. Springer, 2014. 3

[37] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A Dataset for Movie Description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[38] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision (IJCV)*, 123:94–120, 2017. 1, 3, 6, 7

[39] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual Semantic Role Labeling for Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5589–5600, 2021. 3

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 7

[41] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 6

[42] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17959–17968, 2022. 3

[43] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1916–1924, 2017. 3

[44] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *Association of Computational Linguistics (ACL)*, pages 6382–6391, 2019. 3

[45] Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada. Beyond caption to narrative: Video captioning with multiple sentences. In *International Conference on Image Processing (ICIP)*, pages 3364–3368. IEEE, 2016. 3

[46] Soldan, Mattia and Pardo, Alejandro and Alcázar, Juan León and Caba, Fabian and Zhao, Chen and Giancola, Silvio and Ghanem, Bernard. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035, 2022. 3

[47] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV series. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[48] Makarand Tapaswi, Marc T. Law, and Sanja Fidler. Video Face Clustering with Unknown Number of Clusters. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 5

[50] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575. IEEE Computer Society, 2015. 2, 3, 6, 7

[51] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *International Conference on Computer Vision (ICCV)*, pages 4534–4542, 2015. 3

[52] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Compu-*

tational Linguistics: Human Language Technologies, pages 1494–1504, 2015. 3

[53] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7190–7198, 2018. 3

[54] Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1890–1900, 2020. 3

[55] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *International Conference on Computer Vision (ICCV)*, pages 6847–6857, 2021. 3

[56] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10714–10726, 2023. 3

[57] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4584–4593, 2016. 3

[58] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3173, 2017. 3

[59] Youngjae Yu, Jiwan Chung, Jongseok Kim, Heeseung Yun, and Gunhee Kim. LSMDC v2 Challenge presentation. *3rd Workshop on Closing the Loop Between Vision and Language*, 2019. 7

[60] Youngjae Yu, Jongseok Kim, Heeseung Yun, Chung Jiwan, and Gunhee Kim. Character Grounding and Re-Identification inStory of Videos and Text Descriptions. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[61] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020. 3

[62] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8739–8748, 2018. 3