


# Diffusion-based Multimodal Video Captioning

Jaakko Kainulainen, Zixin Guo , and Jorma Laaksonen 

Aalto University, Finland  
kainulainen.jaakko@gmail.com  
{zixin.guo,jorma.laaksonen}@aalto.fi

**Abstract.** Diffusion-based models have recently demonstrated notable success in various generative tasks involving continuous signals, such as image, video, and audio synthesis. However, their applicability to video captioning has not yet received widespread attention, primarily due to the discrete nature of captions and the complexities of conditional generation across multiple modalities. This paper delves into diffusion-based video captioning and experiments with various modality fusion methods and different modality combinations to assess their impact on the quality of generated captions. The novelty of our proposed MM-DiffNet is in the use of diffusion models in multimodal video captioning and in the introduction of a number of mid-fusion techniques for that purpose. Additionally, we propose a new input modality: generated description, which is attended to enhance caption quality. Experiments are conducted on four well-established benchmark datasets, YouCook2, MSR-VTT, VATEX, and VALOR-32K, to evaluate the proposed model and fusion methods. The findings indicate that combining all modalities yields the best captions, but the effect of fusion methods varies across datasets. The performance of our proposed model shows the potential of diffusion-based models in video captioning, paving the way for further exploration and future research in the area.

**Keywords:** Video captioning · Multimodal captioning · Diffusion models · Deep learning

## 1 Introduction

Video captioning involves summarizing a video through generating natural language sentences that describe its context. This complex task bridges computer vision and natural language processing, with the necessity to understand both spatial and temporal aspects of video [21]. The models need to adeptly understand the different modalities and manage to leverage the multimodal information to produce high quality captions [9, 34]. Transformer-based [39] autoregressive models are widely employed for video captioning, typically trained on extensive datasets such as HowTo100M [26] before fine-tuning for specific tasks.

While many video captioning approaches primarily focus on visual features, integrating multimodal aspects, such as speech transcripts and audio, has proven highly beneficial. Speech transcripts generated through automatic speech recognition (ASR) have enhanced the performance of numerous models [3, 7, 18, 23,

35], with this effect being particularly evident in instructional datasets such as YouCook2 [48]. Furthermore, the inclusion of audio provides additional contextual information beyond speech, aiding in the generation of contextually richer captions [2, 3]. Recent research has explored other novel approaches, including integrating knowledge graphs [7] and leveraging large language models (LLMs) [42, 43] for improved caption generation.

In recent years, diffusion models [12, 36] have emerged as powerful tools in generative modeling. They have demonstrated impressive performance across various generative tasks, including image generation [27, 33], audio generation [15, 22], and video generation [11, 14]. However, extending diffusion models to visual captioning remains challenging due to the inherently discrete nature of captions and the complexities of conditional generation across multiple modalities. Recently, the success in diverse generative tasks has inspired a small number of diffusion-based approaches for image captioning [5, 24, 49] and video captioning [38]. Despite achieving promising results, diffusion-based captioning models have not garnered significant attention, with autoregressive models remaining the dominant architecture for captioning tasks.

In this paper, we contribute a novel Multimodal Diffusion Network (MM-Diff-Net) for video captioning and study a new input modality, generated description, to further improve its performance. We evaluate MM-Diff-Net’s behavior on different modality fusion techniques and modality combinations on four datasets, YouCook2, MSR-VTT, VATEX, and VALOR-32K. According to our experiences, the fusion of all available modalities – video, audio, speech transcript and generated description – yields the best video captioning performance with MM-Diff-Net.

## 2 Existing Methods

### 2.1 Video Captioning

Different from image captioning [8, 24], video captioning aims to generate a natural language description summarizing the contents of a given video [21]. This requires the model to understand the multimodal information to produce high-quality descriptions of the video, combining techniques from both computer vision and natural language processing. Successful models recognize objects and activities within frames and understand the video’s temporal progression [32]. Integrating additional modalities, such as audio and speech transcripts, allows the models to generate more informative and contextually rich captions [34].

Captioning models can either be pretrained on extensive datasets and then fine-tuned on specific tasks or trained from scratch using only the benchmark dataset. Although pretraining and finetuning is currently the more common approach, training the models from scratch remains a viable alternative. Older scratch-trained models [10, 21] followed architectures similar to their pretrained counterparts while recent scratch-trained models have explored more novel approaches. TextKG [7] introduced knowledge graphs as an additional modality,

RSFD [47] introduced a frequency-aware diffusion module to enhance the model’s understanding of low-frequency word tokens, and Shen et al. [35] used compressed videos to accelerate data processing while maintaining caption quality.

## 2.2 Modality Fusion

Since models must effectively utilize various video modalities, the method of modality fusion plays a crucial role in their design. Several approaches have been employed in model architectures to address this. Concatenation has been utilized with different modality combinations: AT [10] concatenated video and speech transcript, VALOR [2] concatenated video and audio, and VAST [3] concatenated video, audio, and speech transcript. UniVL [23] implemented hierarchical attention, initially concatenating visual tokens and speech transcripts, which are then fused by an encoder to learn their combined representation. MV-GPT [34] and TextKG [7] employed co-attention, using a fusion encoder with two streams, where each stream performs conditional attention with the other.

## 2.3 Diffusion Models for Captioning

Diffusion models [12, 36] are latent variable models that aim to learn a distribution of data  $x$  by denosing a sample drawn from a normal distribution. Initially crafted for image generation, they have since been adapted to various other generative tasks, such as audio generation [15, 22] and video generation [11, 14]. The diffusion process contains two Markov chains: a forward and a backward process.

**Forward process.** The forward process of the diffusion model is a Markov chain that gradually corrupts the input by adding Gaussian noise at each step. As this process continues, the input becomes increasingly corrupted, until at the end the input is pure random noise. The forward step from the initial state  $x_0$  to a noisier state  $x_t$  is formulated as [37]:

$$x_t = \sqrt{a_t}x_0 + \sqrt{1 - a_t}\epsilon, \quad (1)$$

where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ,  $\alpha_t = 1 - \beta_t$  and  $\beta_t$  is the variance of the added noise at the time step  $t$ . The variance increases at each step, typically following a linear [12] or cosine [28] schedule. Instead of teaching the model to transitions from  $x_t$  to  $x_{t-1}$ , the models are taught to remove all noise from  $x_t$  and directly predict  $x_0$ . This can either be achieved by learning to predict the amount of added noise or by predicting the sample without any noise. The denoising loss is the mean squared error between the caption predicted by the model and the actual caption, formulated as:

$$\mathcal{L} = -\mathbb{E}_{t, x_0, \epsilon} [\|f(x_t, t) - x_0\|^2], \quad (2)$$

where  $f(x_t, t)$  is the model’s estimate of the caption given the noisy input  $x_t$  and the time step  $t$ .

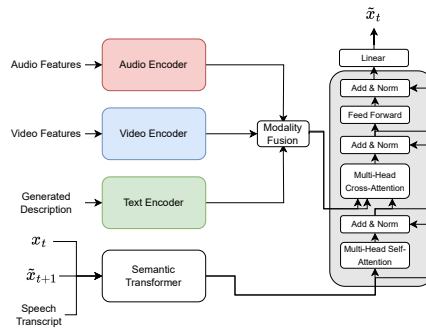
**Reverse process.** During the reverse process, new samples are generated from random noise by gradually removing noise from a noisy sample. The reverse process follows a series of reverse state transitions from sampled noise  $x_T$  to denoised sample  $x_0$ . The diffusion step from  $x_t$  to  $x_{t-1}$  can be expressed as:

$$x_{t-1} = \frac{1}{\sqrt{a_t}} \left( x_t - \frac{\beta_t}{\sqrt{1-a_t}} f(x_t, t) \right) + \sigma_t \epsilon. \quad (3)$$

At each step, the model predicts the noise in the input and removes a portion of it. At the last step,  $\sigma_t \epsilon$  is not added to the output. By iteratively removing noise, the model reconstructs the structure of the data, generating samples that resemble the original data distribution.

**Diffusion-based Captioning.** While diffusion is primarily utilized for tasks involving continuous data spaces, it has also been applied to tasks involving discrete data, such as text generation. Recently, various approaches have been proposed for adapting the diffusion process for generating discrete captions.

DDCap [49] presented an image captioning model that adapts diffusion for discrete data by using masks instead of continuous noise. Bit Diffusion [5] transformed image captions into a continuous form by first converting word tokens into binary bits, which are then represented as real numbers, allowing the model to follow a continuous diffusion process. SCD-Net [24] built on Bit Diffusion by incorporating semantic information from an off-the-shelf retrieval unit, which guides the model during each reverse step to enhance caption quality. CoDi [38] is an any-to-any modality generation model, capable of producing various combinations of video, audio, images, and text by leveraging any mix of these modalities as input. By learning the probability distribution of a latent variable, CoDi processes discrete text like other modalities. Although captioning is not its primary focus, the model achieved competitive results in both image and video captioning. However, it relies solely on visual aspects for video captioning, leaving the multimodal video captioning with a diffusion-based model unexplored.



**Fig. 1:** Architecture of our proposed MM-Diff-Net model.

### 3 Our Method

#### 3.1 Multimodal-Diffusion-Network (MM-Diff-Net)

Figure 1 presents our proposed Multimodal-Diffusion-Network (MM-Diff-Net) model. The architecture follows typical transformer-based encoder-decoder structure, consisting of a visual encoder, an audio encoder, a text encoder, a semantic transformer, and a sentence decoder. This design accommodates multimodal inputs, including four distinct types: video, audio, speech transcript, and textual description. Each encoder generates single-modality representations, which are then fused and fed to the decoder to produce the final caption.

To apply continuous diffusion for generating discrete captions, we follow the approach of Bit Diffusion [5]. First, the words in the caption are tokenized and then converted into  $n = \log_2 W$  binary bits, where  $W$  is the vocabulary size. These bits are subsequently transformed into a vector of real numbers in  $\mathbb{R}^n$  with a dimensionality of  $n$ .

The forward state transition from  $x_0$  to  $x_t$  is formulated as [16]:

$$x_t = \sqrt{\text{sigmoid}(\gamma(-t'))}x_0 + \sqrt{\text{sigmoid}(\gamma(t'))}\epsilon, \quad (4)$$

where  $t$  is the current time step,  $\gamma(t')$  is a monotonically increasing function and  $t' = t/T$ . The model is trained to predict  $x_0$  using the denoising loss, Eq. (2).

Unlike autoregressive models, which generate captions one word at a time, the decoder in a diffusion-based captioning model generates the entire caption at each step of the diffusion process. The reverse process involves a series of diffusion steps, starting from sampled noise  $x_T$  and gradually refining it to the denoised sample  $x_0$ . At each step, the predicted caption is used to estimate and remove a portion of the noise in the input. Following SCD-Net [24] the diffusion step from  $x_t$  to  $x_{t-1}$  is formulated as:

$$x_{t-1} = a_s \left( x_t - \frac{1-c}{a_t} + cf(x_t, \text{sigmoid}(\gamma(t'))) \right) + \sigma\epsilon, \quad (5)$$

where  $s = t - 1 - \Delta$ ,  $\Delta$  is the time difference between steps,  $s' = s/T$ ,  $a_s = \sqrt{\text{sigmoid}(\gamma(-s'))}$ ,  $a_t = \sqrt{\text{sigmoid}(\gamma(-t'))}$ ,  $c = -\text{expm1}(\gamma(-s') - \gamma(-t'))$ ,  $\sigma^2 = \text{sigmoid}(\gamma(s'))c$ , and  $\text{expm1}(\cdot) = \exp(\cdot) - 1$ . After the noise removal process, the model generates the final caption, which is quantized back into bits for the discrete output.

Similar to SCD-Net [24], we add a semantic transformer which brings semantic information to the current latent state  $x_t$  by encoding it with the semantic prior at every diffusion step. However, rather than using a retrieved sentence as the semantic prior, we leverage the speech transcript. Since the speech transcript typically contains language that directly relates to the video content, we use it as the semantic prior to guide the model to produce captions that are better aligned with the video. Additionally, we utilize self-conditioning [5], where the prediction from the previous time step  $\tilde{x}_{t+1}$  is combined with the current latent state

$x_t$ , providing the model with additional information. The semantic-conditional latent state is formulated as:

$$z^x = \mathbf{FC}(\mathbf{Concat}(x_t, \tilde{x}_{t+1})) + \varphi(\gamma(t')) \quad (6)$$

$$z^r = \mathbf{FC}(s_r) \quad (7)$$

$$S^0 = \mathbf{Concat}(z^x, z^r), \quad (8)$$

where  $\varphi(\cdot)$  is a multilayer perceptron, and  $\tilde{x}_{t+1}$  is the prediction of the previous time step. The semantic-conditional latent state is fed to the semantic transformer consisting of  $N$  transformer blocks to generate  $S^N$ , which can be represented as  $S^N = [S_x^N, S_r^N]$ , where  $S_x^N$  is the semantically-conditional latent state that is fed to the decoder, and  $S_r^N$  is ignored.

### 3.2 Modalities

MM-Diff-Net can utilize features from four different modalities in caption generation: video (V), audio (A), speech transcript (T) and generated description (G). The generated description modality is an additional modality that we propose, created by leveraging an LLM-based video understanding model, NExT-GPT [43]. This approach aims to improve the captioning process by providing a short description of the video content grounded in its visual elements. The NExT-GPT model is tasked to analyze the visual content of the videos and to generate short descriptions consisting of a few sentences that provide a general overview of the video content. Combined with the other more traditional modalities, the description contributes in the model generating higher quality captions. However, the descriptions frequently contain hallucinations, such as noise or irrelevant information. Therefore, it is imperative for the model to learn to identify the relevant parts of the description and filter out the irrelevant information.

### 3.3 Fusion Methods

We have devised and experimented with five different modality fusion techniques for MM-Diff-Net. The first method explored is **concatenation fusion**:

$$C = \mathbf{Concat}(V, A, G), \quad (9)$$

where V, A and G are the visual, audio and generated description tokens, respectively. Somewhat more versatile fusion can be obtained by following the architecture of UniVL [23] and using **hierarchical attention fusion**:

$$C = \mathbf{Transformer}(\mathbf{Concat}(V, A, G)). \quad (10)$$

The last three methods utilize cross-attention in various ways, enabling the model to simultaneously attend to multiple modalities and combine information from one modality with another. Unlike the self-attention, which attends to different positions within the same modality, cross-attention attends to positions

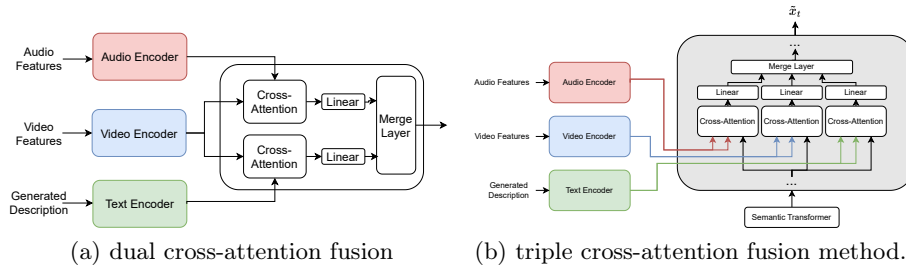


Fig. 2: Dual cross-attention and triple cross-attention fusion methods.

in one modality based on the information in another modality. This alignment allows the model to synchronize the representations of the different modalities effectively. The first one of these methods, **cross-attention fusion**, concatenates audio and text embeddings and merges them with visual embeddings using cross-attention. In this layer, the video embedding serves as the query, while the combined audio and text embedding act as the keys and values. This method represents a straightforward approach to leveraging cross-attention for combining three modalities and can be formulated as:

$$E = \text{Concat}(A, G) \tag{11}$$

$$V_{k+1} = \text{MultiHead}(V_k, E, E), \tag{12}$$

where  $V_{k+1}$  is the output of the  $k$ -th cross-attention layer and  $V_0$  is the output of the visual encoder.

The fourth method, **dual cross-attention fusion**, shown in Figure 2(a), was inspired by [19]. Similar to the previous method, audio and text embeddings are combined with the video embeddings using cross-attention. However, instead of merging audio and text embeddings before combining them with video embeddings, this method combines them with video embeddings separately. An additional cross-attention layer is added alongside the existing one in the cross-attention block. Consequently, the transformer has two adjacent cross-attention layers: one for combining audio with video and another for combining text with video. This dual setup enables the separate integration of audio and text with video, increasing the model’s understanding of video-audio and video-text relationships. The outputs of the cross-attention layers are then processed through linear layers, and subsequently combined using a merge layer. The method can be formulated as:

$$V_k^m = \text{FC}(\text{MultiHead}(V_k, m, m)), \quad m = A, G \tag{13}$$

$$V_{k+1} = \text{FC}(\text{Concat}(V_k^A, V_k^G)), \tag{14}$$

where  $m$  are the embeddings of the corresponding modality.

The last method, **triple cross-attention fusion**, shown in Figure 2(b), keeps the video, audio and text embeddings separate. It combines them individ-

ually with the decoder input using cross-attention. To achieve this, two additional cross-attention layers are introduced to the decoder, positioned between the self-attention and feed-forward layers. Triple cross-attention consists of the three adjacent cross-attention layers, three linear layers and a merge layer. Each cross-attention layer uses the outputs of the self-attention layer as the query and one of video, audio or text embeddings as the value and key. This enables the model to learn the contributions of each modality to the caption creation with pairwise connections between the modalities and the caption as:

$$h_i^m = \mathbf{FC}(\mathbf{MultiHead}(h_i, m, m)), \quad m = V, A, G \quad (15)$$

$$\tilde{h}_i = \mathbf{FC}(\mathbf{Concat}(h_i^V, h_i^A, h_i^G)), \quad (16)$$

where  $h_i$  is the output of the  $i$ -th self-attention layer and  $\tilde{h}_i$  is the output of the  $i$ -th merge layer.

## 4 Experiments and Results

### 4.1 Experimental Settings

We conduct experiments on the YouCook2 [48], MSR-VTT [44], VATEX [41] and VALOR-32K [2] datasets. Our captioning models are evaluated on BLEU@4 (B) [29], METEOR (M) [1], ROUGE-L (R) [20] and CIDEr-D (C) [40].

Following UniVL [23], video features pre-extracted by S3D [45] are used for both the YouCook2 and MSR-VTT datasets. The video features for VATEX and VALOR-32K were extracted using the pre-trained CLIP-ViT-L/14 model [31]. The automatic speech recognition (ASR) transcript for the YouCook2 dataset was provided by UniVL [23], and the ASR transcripts for the MSR-VTT, VATEX and VALOR-32K datasets were generated specifically for these experiments using the Azure AI speech-to-text service [25]. Audio features for all datasets were extracted using the pretrained ImageBind model [6], and the generated descriptions were created using the pretrained NExT-GPT model [43] with the prompt “Describe this video.”.

Each encoder and decoder consists of three layers, with a hidden layer size of  $d = 512$  and  $h = 8$  heads. Optimization is performed using mean squared error and label smoothing, utilizing the Adam optimizer [17], with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . During training, an exponential moving average of the weights is used. The model is trained for 60 epochs using the same learning rate schedule as in [39], with 2,000 warm-up steps for YouCook2, 20,000 warm-up steps for MSR-VTT, and 8,000 warm-up steps for VATEX and VALOR-32K, respectively. Inference time depends largely on the number of diffusion steps, so to balance inference time and performance, the model uses  $T = 50$  time steps during inference. The maximum length of the predicted captions is 20 words. The batch sizes are set to 64 for YouCook2, VATEX, and VALOR-32K and 128 for MSR-VTT.

The vocabulary used in the experiments was constructed from the captions, ASR transcripts, and generated descriptions within the training sets. The rare



words were removed by selecting words that occur at least four times. The YouCook2 vocabulary contains  $W = 4,702$  words, requiring 13 bits to represent each word, the MSR-VTT vocabulary contains  $W = 10,738$  words, requiring 14 bits to represent each word, the VATEX vocabulary contains  $W = 13,261$  words, requiring 14 bits to represent each word and the VALOR-32K vocabulary contains  $W = 7,988$  words, requiring 13 bits to represent each word.

## 4.2 Different Fusion Methods

Table 1 presents the captioning results obtained using the different fusion techniques introduced in Section 3.3 when all available input modalities were used. For the YouCook2 dataset, the results indicate the significance of cross-attention-based fusion methods. Particularly the last two methods, dual cross-attention and triple cross-attention, emerge as the most effective. These methods show comparable BLEU-4, METEOR, and ROUGE-L scores, with triple cross-attention notably outperforming others in terms of CIDEr-D. Since the relative effectiveness of the different modalities varies significantly, the results suggest that keeping them separate for as long as possible yields the best performance for YouCook2. Concatenation and hierarchical attention fusion methods generate inferior results compared to cross-attention-based methods. Given the minimal impact of audio and generated description modalities on captioning results (to be seen in Section 4.3), combining them with video tokens through concatenation appears to diminish the effectiveness of the video modality.

For the MSR-VTT dataset, the performance across all fusion methods remains relatively consistent. Among these methods, hierarchical attention stands out, delivering particularly strong results in terms of CIDEr-D, while also maintaining competitive performance across the other metrics. Given the relatively high effectiveness of video, audio, and generated description individually (in Section 4.3), concatenation-based methods demonstrate better performance compared to those observed with the YouCook2 dataset.

For VATEX, the hierarchical attention delivers the strongest performance, followed by the triple-cross-attention method. These results suggest that both early and late fusion strategies can be effective for VATEX. However, concatenation performs worse than hierarchical fusion, indicating that VATEX benefits from slightly more complex fusion techniques.

In contrast, for VALOR-32K, concatenation and hierarchical attention fusion methods outperform cross-attention-based methods, with simple concatenation achieving the best results. This suggests that VALOR-32K benefits from early fusion without the need for complex techniques.

## 4.3 Different Modality Combinations

Next, we experimented with various modality combinations as shown in the results of Table 2. Following the findings of Table 1, triple cross-attention fusion method, Eqs. (15) and (16), is used with YouCook2, hierarchical attention fusion

**Table 1:** Different fusion methods on the YouCook2, MSR-VTT, VALOR-32K and VATEX test splits.

Fusion method	YouCook2				MSR-VTT				VATEX				VALOR-32K			
	B	M	R	C	B	M	R	C	B	M	R	C	B	M	R	C
concatenation	9.11	38.90	16.67	113.78	<u>44.29</u>	62.06	27.80	47.05	36.17	24.14	52.08	57.69	<b>5.09</b>	<b>11.31</b>	<b>26.46</b>	<b>28.94</b>
hierarchical attention	9.08	39.00	<u>16.88</u>	114.18	44.07	<u>62.17</u>	<b>28.03</b>	<b>48.86</b>	<b>36.97</b>	<b>24.32</b>	<b>52.46</b>	<b>58.93</b>	<u>4.91</u>	<u>11.12</u>	26.08	<u>28.61</u>
cross-attention	9.18	<u>39.26</u>	16.64	115.41	44.00	<b>62.38</b>	<u>27.98</u>	47.61	36.24	24.02	51.88	57.09	4.50	10.62	25.82	26.93
dual cross-attention	<b>9.68</b>	39.22	<b>16.93</b>	115.61	43.97	61.90	27.93	<u>47.82</u>	<u>36.88</u>	24.18	52.20	58.13	4.85	10.78	<u>26.34</u>	27.64
triple cross-attention	<u>9.47</u>	<b>39.46</b>	16.87	<b>117.24</b>	<b>44.72</b>	61.89	27.82	46.89	36.87	<u>24.21</u>	<u>52.23</u>	<u>58.51</u>	4.54	10.39	25.98	26.34

method, Eq. (10), is used with MSR-VTT and VATEX, whereas concatenation, Eq. (9), is used with VALOR-32K, when fusion methods are required.

The results reveal that, across all datasets, the visual input emerges as the strongest single modality. In YouCook2, the speech transcript also demonstrates significant strength, aligning with the instructional nature of the videos where verbal explanations are closely tied to the cooking process. On the contrary, audio and generated description offer limited assistance in YouCook2, due to the presence of ambient noise and the specific nature of the captions focused on ingredients and cooking techniques. Combining video and speech transcript enhances results compared to individual modalities, while adding audio or description provides only marginal improvements.

In contrast, the speech transcript proves less effective in MSR-VTT, with minimal improvements observed when combined with other modalities. Both audio and description contain relevant information, with description yielding slightly better results. When combined with visual input, audio and description offer similar improvements, suggesting that audio provides additional information not captured by the visual modality, such as background music.

In both VATEX and VALOR-32K, both speech transcript and audio produce poor results individually, with audio slightly outperforming speech transcript, while the generated description produces better results. However, when combined with the visual input, the impact of these modalities varies between these two datasets. In VATEX, the performances of all three modalities show similar improvements, whereas in VALOR-32K the generated description improves the results more than either audio or speech transcript.

Optimal performance is achieved with all modalities combined in all datasets, indicating the robustness of the modality selection across the datasets. However, the second-best results are obtained with different combinations: "VTG" for YouCook2 and "VAG" for MSR-VTT, VATEX and VALOR-32K. Interestingly, the second-best results are nearly as good as the best ones, indicating that the left-out modalities have a minimal impact on performance. This aligns with the observation that these left-out modalities are the weakest individually for each dataset.

#### 4.4 State-of-the-Art Comparison

The comparison results with the state-of-the-art models for YouCook2, MSR-VTT, VATEX and VALOR-32K are presented in Table 3. The results for MM-

**Table 2:** Captioning results on the YouCook2, MSR-VTT, VATEX and VALOR-32K datasets with different modality combinations. V, A, T and G stand for video, audio, speech transcript and generated description, respectively.

Modality	YouCook2				MSR-VTT				VATEX				VALOR-32K			
	B	M	R	C	B	M	R	C	B	M	R	C	B	M	R	C
V	7.41	36.06	14.40	98.73	39.89	60.07	27.10	45.34	35.2	23.58	51.47	54.75	4.66	10.65	25.97	26.42
A	1.61	22.40	6.86	18.14	36.05	55.22	22.90	26.00	12.59	13.11	36.98	5.79	3.60	8.54	23.52	9.21
T	6.99	32.86	13.92	81.13	25.13	49.45	18.41	12.66	10.64	12.84	36.56	4.86	2.07	7.52	22.35	8.12
G	1.80	23.87	7.46	23.75	33.01	54.93	22.57	29.85	21.04	16.98	41.90	21.22	4.21	9.32	24.09	16.23
VA	7.84	36.46	14.78	101.19	<b>44.50</b>	61.68	27.90	46.74	35.84	23.79	52.12	55.95	4.99	11.00	26.08	27.20
VT	9.18	38.39	16.71	114.00	41.24	60.08	26.60	45.43	36.27	23.95	52.03	56.09	4.91	10.81	25.96	27.03
VG	7.50	36.96	14.62	101.37	43.74	61.59	27.77	46.59	36.17	23.80	52.09	56.02	5.05	10.88	26.12	28.05
TG	7.25	33.85	14.19	86.60	35.40	55.94	23.69	30.68	22.94	17.61	43.02	23.05	3.71	8.95	23.42	16.72
VAT	8.92	38.58	16.30	115.00	<u>44.37</u>	61.64	27.77	46.68	36.65	24.02	52.27	57.22	5.02	11.01	26.25	28.23
VAG	7.80	36.72	14.75	101.39	43.92	<u>61.83</u>	<u>28.02</u>	<u>48.73</u>	<u>36.71</u>	<u>24.16</u>	<u>52.28</u>	<u>58.41</u>	<u>5.05</u>	<u>11.14</u>	26.12	<u>28.46</u>
VTG	<b>9.56</b>	<u>39.42</u>	<u>16.87</u>	<u>116.54</u>	43.03	61.43	27.51	46.88	36.03	24.01	51.92	56.91	5.04	10.99	<u>26.24</u>	28.38
VATG	<u>9.47</u>	<b>39.46</b>	<b>16.88</b>	<b>117.24</b>	44.07	<b>62.17</b>	<b>28.03</b>	<b>48.86</b>	<b>36.97</b>	<b>24.32</b>	<b>52.46</b>	<b>58.93</b>	<b>5.09</b>	<b>11.31</b>	<b>26.46</b>	<b>28.94</b>

Diff-Net were obtained using two stacked diffusion models in a cascaded fashion, as detailed in Section 4.5. This approach requires training the models twice and was employed exclusively for these results. The tables display two sets of results for MM-Diff-Net: one with modalities comparable to other models and another with all modalities.

For YouCook2, TextKG [7] demonstrates the strongest performance across three metrics: METEOR, ROUGE-L, and CIDEr-D, while MV-GPT performs best for BLEU-4. MM-Diff-Net struggles to match TextKG’s performance, indicating that the addition of knowledge graphs significantly improves results. However, MM-Diff-Net achieves comparable results with "VT" modalities compared to other methods, and "VATG" modality combination secures the second-best performance for both METEOR and CIDEr-D.

**Table 3:** Performance comparison with the state-of-the-art models trained from scratch on the YouCook2, MSR-VTT, VATEX and VALOR-32K datasets. Triple cross-attention fusion was used with YouCook2, concatenation with VALOR-32K and hierarchical attention fusion with MSR-VTT and VATEX. Two stacked diffusion models were used for all datasets.

YouCook2					
Model	Modality	B	M	R	C
SwinBERT [21]	V	9.00	37.30	15.60	109.00
AT [10]	VT	9.00	36.70	<u>17.80</u>	112.00
UniVL [23]	VT	9.46	37.44	16.27	115.00
MV-GPT [34]	VT	<b>13.25</b>	35.48	17.56	103.00
TextKG [7]	VT	<u>11.70</u>	<b>40.20</b>	<b>18.40</b>	<b>133.00</b>
MM-Diff-Net (ours)	VT	9.73	39.31	17.02	116.57
MM-Diff-Net (ours)	VATG	9.92	<u>39.85</u>	17.19	<u>120.60</u>

MSR-VTT					
Model	Modality	B	M	R	C
SwinBERT [21]	V	41.90	62.10	<u>29.90</u>	<u>53.80</u>
RSFD [47]	V	43.40	62.20	29.30	53.10
TextKG [7]	V	43.70	<u>62.40</u>	29.60	52.40
Shen et al. [35]	V	<b>44.40</b>	<b>63.40</b>	<b>30.30</b>	<b>57.20</b>
MM-Diff-Net (ours)	V	41.37	60.45	27.11	46.18
MM-Diff-Net (ours)	VATG	<u>43.72</u>	62.04	28.31	49.55

VATEX					
Model	Modality	B	M	R	C
Support-set [30]	V	32.80	24.40	49.10	51.20
OpenBook [46]	V	33.90	23.70	50.20	57.50
Shen et al. [35]	V	35.80	<u>25.30</u>	52.00	<u>64.80</u>
SwinBERT [21]	V	<b>38.70</b>	<b>26.20</b>	<b>53.20</b>	<b>73.00</b>
MM-Diff-Net (ours)	V	36.22	23.86	52.08	55.56
MM-Diff-Net (ours)	VATG	<u>37.28</u>	24.44	<u>52.58</u>	59.15

VALOR-32K					
Model	Modality	B	M	R	C
SwinBERT [21]	V	<u>5.40</u>	10.70	<u>27.20</u>	27.30
SMPFF [4]	VA	<b>7.50</b>	<b>12.60</b>	<b>28.60</b>	<u>37.10</u>
VAST [3]	VAT	-	-	-	<b>40.80</b>
MM-Diff-Net (ours)	V	5.12	11.01	26.44	28.23
MM-Diff-Net (ours)	VATG	5.19	<u>11.51</u>	27.00	31.53

In contrast, for MSR-VTT, MM-Diff-Net falls notably behind the state-of-the-art methods. This is particularly evident when only the visual modality is used. Even when all modalities are utilized, MM-Diff-Net achieves only the second-best BLEU-4 score, with other metrics still trailing behind the state-of-the-art models. Shen et al. [35] demonstrates markedly superior performance across all metrics compared to all the other models.

For VATEX, SwinBERT [21] significantly outperforms all the other models across all metrics. For METEOR and CIDEr-D metrics the performance of MM-Diff-Net falls behind the performance of Shen et al. [35]. However, although MM-Diff-Net struggles to match the performance of SwinBERT, it achieves the second-best results for both the BLEU-4 and ROUGE-L metrics.

For VALOR-32K, VAST [3] reports the strongest performance on CIDEr-D, but does not report the other metrics, while SMPFF [4] performs best on the other metrics. MM-Diff-Net fails to match the performance of these models, but with only the visual modality it is on par with SwinBERT [21] and outperforms it when using all the modalities.

#### 4.5 Ablation Studies

Comprehensive ablation studies were conducted to analyze the effectiveness of the model design and the impact of different parameters and architectural changes. The ablation studies on the YouCook2 dataset were performed using the triple cross-attention method, Eqs. (15) and (16), and the test split of the dataset. Conversely, the ablation studies on the MSR-VTT dataset utilized the hierarchical attention method, Eq. (10), and the validation split.

**Cascaded diffusion models.** We examined the effect of stacking multiple diffusion models [13,24], which aim to strengthen the sentence decoder output, in a cascaded fashion. Model  $f_i$  ( $i \geq 2$ ) is conditioned using the predicted sentence  $x_0^{i-1}$  from the previous diffusion model. This, combined with the latent state  $x_t$  and the prediction of the previous time step  $\tilde{x}_{t+1}$ , leads to Eq. (6) being reshaped:

$$z^{x,i} = \mathbf{FC}(\mathbf{Concat}(x_t, \tilde{x}_{t+1}, x_0^{i-1})) + \varphi(y(t')). \quad (17)$$

The effectiveness of this cascading approach is evaluated by comparing the performance of using two or three stacked diffusion models against using only one diffusion model. The results presented in Table 4 demonstrate a noticeable improvement when employing two stacked diffusion models compared to using just one. However, further stacking beyond two models does not yield a significant improvement in performance.

**Table 4:** Stacking different numbers of diffusion models on MSR-VTT.

Diffusion models	B	M	R	C
1	45.05	62.36	28.42	49.97
2	<u>46.09</u>	<u>63.32</u>	<b>29.16</b>	<u>51.03</u>
3	<b>46.12</b>	<b>63.37</b>	<u>29.12</u>	<b>51.06</b>

**Table 5:** Different numbers of transformer layers on MSR-VTT.

Layers	B	M	R	C
2	<b>45.12</b>	61.81	<u>28.50</u>	49.17
3	<u>45.05</u>	<b>62.36</b>	28.42	<u>49.97</u>
4	44.98	62.23	28.28	49.89
5	44.95	<u>62.33</u>	<b>28.57</b>	<b>50.07</b>

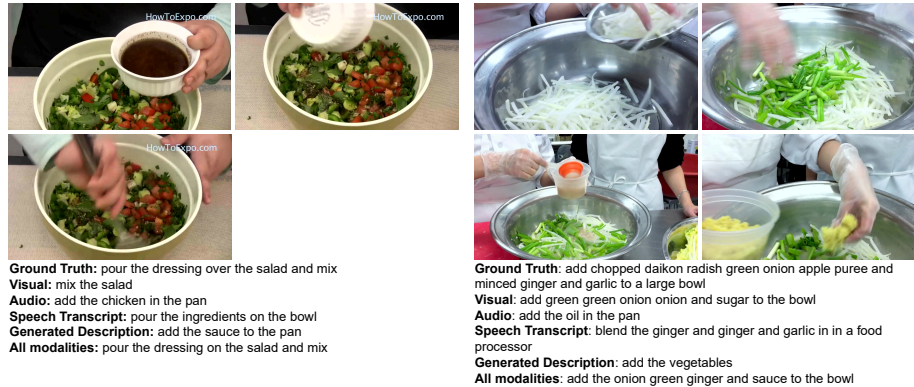
**Number of transformer layers.** The impact of varying the number of transformer layers is examined next, with the results presented in Table 5. These results indicate that increasing the number of layers does not significantly affect the outcomes. Notably, METEOR and CIDEr-D scores improve when the number of transformer layers is increased from two to three, but further increases have a smaller impact, while other metrics remain relatively unchanged. Consequently, the number of transformer layers was set to three, balancing performance and computational efficiency effectively.

**Switching semantic prior.** Since the semantic prior has a large role in the model architecture, as discussed in Section 3.1, the effect of using the generated description as the semantic prior is tested. In our experiment, there is no noticeable difference in results with MSR-VTT as both the BLEU-4 and CIDEr-D measures are practically equal when using either the generated description or the speech transcript as the semantic prior. On the other hand, changing the speech transcript as the prior to generated description leads to a significant decrease in performance with YouCook2 as BLEU-4 drops from 9.47 to 8.59 and CIDEr-D from 117.24 to 113.35.

With MSR-VTT, either of the text-based modalities can thus serve as a semantic prior, but in YouCook2, the speech transcript should be used. Since the videos in YouCook2 dataset are instructional, the speech transcript is more likely to contain words relevant to the caption. Therefore the generated description may contain extraneous or irrelevant information, potentially disrupting the captioning process.

#### 4.6 Qualitative Results

Figure 3 shows sample captions generated for two videos of the YouCook2 dataset employing various modalities. The first sample shows a case where the MM-DiffNet model has successfully generated a coherent, high-quality caption, whereas the second one illustrates the difficulties the model may encounter. In the examples, the influence of different modalities is evident. Captions generated using the audio elements are entirely unrelated to the video, while description-based captions struggle to capture precise details in the cooking process. Captions created using the visual elements and speech transcripts capture distinct elements of the video, with each providing a partial answer. Combining all modalities results in



**Fig. 3:** Qualitative results generated by using the different modalities for two videos from the YouCook2 dataset.

higher-quality captions compared to any single modality. The caption for the first video closely resembles the ground truth, while the one for the second video captures the main action, but fails to recognize all ingredients accurately.

## 5 Conclusions and Future Work

In this paper, we addressed the multimodal video captioning task by proposing a novel Multimodal-Diffusion-Network (MM-Diff-Net) that employs a diffusion process to generate multimodally grounded captions. Diffusion models have faced success in many recent research tasks, but to our best knowledge, our work is the first one to apply them in multimodal video captioning.

Our proposed MM-Diff-Net model obtained competitive results on all four datasets, YouCook2, MSR-VTT, VATEX and VALOR-32K, compared to the performance of the best state-of-the-art autoregressive models. It is interesting to note that for each dataset, a different SOTA model performed the best. MM-Diff-Net’s ability to generate coherent captions is thus noteworthy and promising. Also, we found that the inclusion of our proposed new input modality, generated description, consistently improved the quality of the generated video captions. In experiments on four datasets the fusion of all available modalities, video, audio, speech transcript and generated description, yielded the best performance. However, the optimal modality fusion method varied across the datasets. A topic for future research would thus be to study this behavior in depth and develop a domain-agnostic fusion model that could be optimal for all datasets.

**Acknowledgments.** This study was funded by Research Council of Finland (USSEE project, grant number 345791) and computational resources were provided by LUMI supercomputer, owned by the EuroHPC JU, hosted by CSC.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
2. Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., Liu, J.: VALOR: Vision-audio-language omni-perception pretraining model and dataset. arXiv preprint arXiv:2304.08345 (2023)
3. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 36 (2024)
4. Chen, S., Zhu, X., Hao, D., Liu, W., Liu, J., Zhao, Z., Guo, L., Liu, J.: MM21 Pre-training for video understanding challenge: Video captioning with pretraining techniques. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4853–4857 (2021)
5. Chen, T., Zhang, R., Hinton, G.: Analog bits: Generating discrete data using diffusion models with self-conditioning. In: Proceedings of the Eleventh International Conference on Learning Representations (2022)
6. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (2023)
7. Gu, X., Chen, G., Wang, Y., Zhang, L., Luo, T., Wen, L.: Text with knowledge graph augmented transformer for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18941–18951 (June 2023)
8. Guo, Z., Wang, T.J., Laaksonen, J.: Clip4idc: Clip for image difference captioning. In: Proceedings of the Asia-Pacific Chapter of the Association for Computational Linguistics. pp. 33–42 (2022)
9. Guo, Z., Wang, T.J.J., Laaksonen, J.: Post-attention modulator for dense video captioning. In: International Conference on Pattern Recognition (ICPR). pp. 1536–1542. IEEE (2022)
10. Hessel, J., Pang, B., Zhu, Z., Soricut, R.: A case study on combining asr and visual features for generating instructional video captions. arXiv preprint arXiv:1910.02930 (2019)
11. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
13. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research* **23**(47) (2022)
14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 35, pp. 8633–8646 (2022)
15. Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion

- models. In: Proceedings of the International Conference on Machine Learning. pp. 13916–13932. PMLR (2023)
16. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Proceedings of the Advances in neural information processing systems **34**, 21696–21707 (2021)
  17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
  18. Ko, D., Choi, J., Choi, H.K., On, K.W., Roh, B., Kim, H.J.: MELTR: Meta loss transformer for learning to fine-tune video foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20105–20115 (2023)
  19. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the International Conference on Machine Learning. pp. 12888–12900 (2022)
  20. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. pp. 605–612 (2004)
  21. Lin, K., Li, L., Lin, C.C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., Wang, L.: SwinBERT: End-to-end transformers with sparse attention for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17949–17958 (2022)
  22. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503 (2023)
  23. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
  24. Luo, J., Li, Y., Pan, Y., Yao, T., Feng, J., Chao, H., Mei, T.: Semantic-conditional diffusion networks for image captioning\*. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23359–23368 (2023)
  25. Microsoft Corporation: Azure AI Speech to Text, <https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/>
  26. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2630–2640 (2019)
  27. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
  28. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the International Conference on Machine Learning. pp. 8162–8171 (2021)
  29. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
  30. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020)



31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
32. Radman, A., Laaksonen, J.: As-net: active speaker detection using deep audio-visual attention. *Multimedia Tools and Applications* pp. 1–16 (2024)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
34. Seo, P.H., Nagrani, A., Arnab, A., Schmid, C.: End-to-end generative pretraining for multimodal video captioning. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17938–17947 (2022)
35. Shen, Y., Gu, X., Xu, K., Fan, H., Wen, L., Zhang, L.: Accurate and fast compressed video captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15558–15567 (2023)
36. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the International Conference on Machine Learning. pp. 2256–2265 (2015)
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
38. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 36 (2024)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 30 (2017)
40. Vedantam, R., Zitnick, C., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015)
41. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatec: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4581–4591 (2019)
42. Wang, Z., Wang, L., Zhao, Z., Wu, M., Lyu, C., Li, H., Cai, D., Zhou, L., Shi, S., Tu, Z.: GPT4Video: A unified multimodal large language model for instruction-followed understanding and safety-aware generation. arXiv preprint arXiv:2311.16511 (2023)
43. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: NExT-GPT: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023)
44. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5288–5296 (2016)
45. Zhang, D., Dai, X., Wang, X., Wang, Y.F.: S3D: single shot multi-span detector via fully 3D convolutional networks. arXiv preprint arXiv:1807.08069 (2018)
46. Zhang, Z., Qi, Z., Yuan, C., Shan, Y., Li, B., Deng, Y., Hu, W.: Open-book video captioning with retrieve-copy-generate network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9837–9846 (2021)

47. Zhong, X., Li, Z., Chen, S., Jiang, K., Chen, C., Ye, M.: Refined semantic enhancement towards frequency diffusion for video captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3724–3732 (2023)
48. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7590–7598 (2018)
49. Zhu, Z., Wei, Y., Wang, J., Gan, Z., Zhang, Z., Wang, L., Hua, G., Wang, L., Liu, Z., Hu, H.: Exploring discrete diffusion models for image captioning. arXiv preprint arXiv:2211.11694 (2022)