Diff-LMM: Diffusion Teacher-Guided Spatio-Temporal Perception for Video Large Multimodal Models

Jisheng Dang[1,2,3], **Ligen Chen**[1], **Jingze Wu**[1], **Ronghao Lin**[1], **Bimei Wang**[4,3], **Yun Wang**[5], **Liting Wang**[6], **Nannan Zhu**[1] and **Teng Wang**[7]
[1] Sun Yat-sen University, Guangdong, China
[2] Lanzhou University, Gansu, China
[3] National University of Singapore, Singapore
[4] Jinan University, Guangdong, China
[5] City University of Hong Kong, China
[6] Northwest Normal University, Gansu, China
[7] University of Hong Kong, China
dangjsh@mail2.sysu.edu.cn, chenlg8@mail2.sysu.edu.cn, tengwang@connect.hku.hk

Abstract

Dynamic spatio-temporal understanding is essential for video-based multimodal tasks, yet existing methods often struggle to capture fine-grained temporal and spatial relationships in long videos. Current approaches primarily rely on pre-trained CLIP encoders, which excel in semantic understanding but lack spatially-aware visual context. This leads to hallucinated results when interpreting fine-grained objects or scenes. To address these limitations, we propose a novel framework that integrates diffusion models into multimodal video models. By employing diffusion encoders at intermediate layers, we enhance visual representations through feature alignment and knowledge distillation losses, significantly improving the model's ability to capture spatial patterns over time. Additionally, we introduce a multi-level alignment strategy to learn robust feature correspondence from pre-trained diffusion models. Extensive experiments on benchmark datasets demonstrate our approach's state-of-the-art performance across multiple video understanding tasks. These results establish diffusion models as a powerful tool for enhancing multimodal video models in complex, dynamic scenarios.

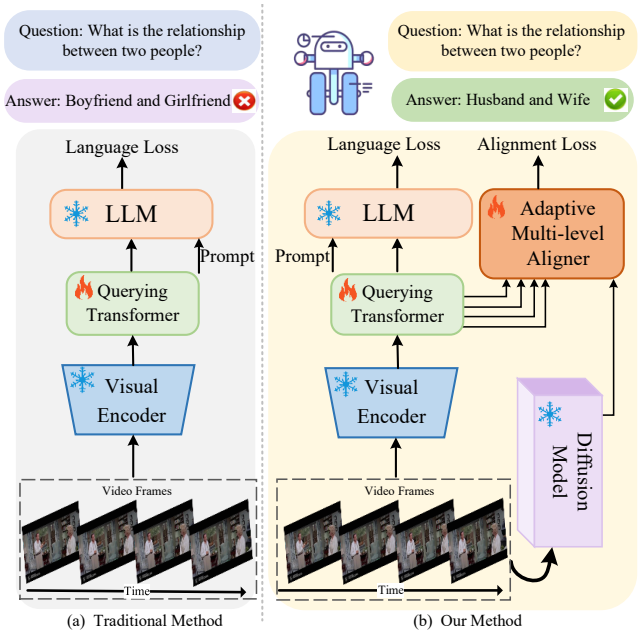

Figure 1: **Traditional method vs. the proposed Diff-LMM**. Previous methods (a) employ the Querying Transformer (Q-Former) to connect the visual encoder with the LLM decoder, guided by language generation loss. In contrast, our approach (b) introduces diffusion-based supervision, enabling the Q-Former to capture fine-grained cues and enhance spatio-temporal dynamics.

1 Introduction

Recently, multimodal large language models (MLLMs) [Zhang *et al.*, 2023; Li *et al.*, 2023a] have demonstrated significant advancements in visual understanding tasks, including image or video recognition [Li *et al.*, 2024b; Ma *et al.*, 2024; Meng *et al.*, 2024; Wang *et al.*, 2024b; Meng *et al.*, 2025], visual question-answering [Li *et al.*, 2023a] and object segmentation [Dang *et al.*, 2023b; Dang *et al.*, 2024a; Dang *et al.*, 2024c]. However, the application of MLLMs in long video understanding remains challenging due to the complexity of modeling spatio-temporal dependencies across consecutive video frames. These models must effectively reason about complex dynamics, intricate scenes, and subtle visual details over extended periods, akin to human information extraction from complex visual streams. Consequently, developing efficient and effective solutions for this task remains a significant challenge, which limits its application in time-sensitive fields [Dang and Yang, 2021; Dang and Yang, 2022].

Existing works in video-based multimodal tasks can be divided into two categories. The first category includes CLIP-based methods (e.g., Video-LLaMA [Zhang *et al.*, 2023], MA-LMM [He *et al.*, 2024], and TimeChat [Ren *et al.*, 2024]), which extract features using CLIP and convert them to the LLM embedding space via multi-level perception (MLP) or Q-Former. These methods prioritize semantic understanding but overlook spatial context, which limits their ability to model spatio-temporal relationships. The second category involves self-supervised backbones(e.g., DINO [Oquab *et al.*, 2023]), or combinations of CLIP and self-supervised encoders [Zong *et al.*, 2024]. These approaches capture more visual cues, balancing coarse-grained and fine-grained understanding. However, they have limitations: (a) multiple encoders increase computational costs, making them impractical for processing long videos, and (b) the potential of large-scale generative models, such as diffusion models, to improve MLLM performance remains unexplored. How to efficiently and effectively capture the complex spatio-temporal dynamics present in long video sequences remains a challenging problem.

To address the aforementioned challenges, we propose a novel framework, Diff-LMM, which introduces diffusion-guided supervision into large multimodal models. Our solution is simple, efficient, and effective, integrating pre-trained knowledge from diffusion models to enhance fine-grained understanding. As shown in Figure 1, the core idea is using diffusion features as intermediate supervision within the visual encoding layers and employing multi-level distillation loss to achieve feature alignment. Diff-LMM offers several advantages: (a) Instead of directly incorporating the diffusion backbone as the visual encoder, the teacher-student distillation mechanism allows the model to learn rich pre-trained features without increasing inference costs. (b) Intermediate intervention in the visual encoding layers complements missing visual details that are challenging to capture through language supervision alone. (c) By employing multi-level alignment loss, we alleviate the representation space gap between the diffusion model and our encoder, achieving adaptive and effective alignments.

Experiments demonstrate that Diff-LMM achieves state-of-the-art performance across multiple long video understanding benchmarks. Ablation analysis further confirms that visual representations from pre-trained diffusion models, such as DiT, offer a positive effect in fine-grained tasks within long video scenarios. These findings underscore the substantial value of diffusion models in enhancing multimodal video models, particularly in complex environments.

We summarize our main contributions as follows:

- We present a novel diffusion-based supervision for MLLMs, utilizing a meticulously crafted feature alignment strategy to enhance fine-grained representations of long videos.
- We introduce an adaptive multi-level alignment mechanism that dynamically adjusts alignment granularity across various levels, effectively bridging the representation gap between teacher and student models and establishing robust inter-model correspondences.
- Extensive experiments demonstrate that leveraging pre-trained diffusion models significantly improves performance on fine-grained tasks in long video scenarios, underscoring the potential of diffusion models as a powerful tool for advancing multimodal video models in complex, dynamic environments.

2 Related Works

2.1 Long Video Understanding

Recent advancements in MLLMs have significantly improved their application to video understanding tasks. To process video inputs, image-language models [Li *et al.*, 2023a] typically flatten spatio-temporal features into one-dimensional sequences and utilize a pre-trained large language model (LLM) for decoding. However, these approaches face challenges in capturing temporal dynamics due to the context length limitations of LLMs and the high GPU consumption required for processing.

To address these challenges, several works [Maaz *et al.*, 2023; Dang *et al.*, 2023a; Dang *et al.*, 2024b] have attempted to uniformly sample from video, aiming to preserve as much information as possible under limited input conditions. For instance, Video-ChatGPT [Maaz *et al.*, 2023] applies average pooling modules to reduce input dimensions, which, however, results in a significant loss of visual details. Other works aim to preserve the maximum number of frames by designing specialized memory modules. Inspired by the Atkinson-Shiffrin memory model, MovieChat [Song *et al.*, 2024] introduces memory modules to retain detailed video content effectively. However, these methods still lack explicit temporal modeling, resulting in suboptimal performance. In contrast, Video-LLaMA [Zhang *et al.*, 2023] utilizes an additional query transformer to directly model temporal relationships, although this increases computational complexity. MA-LMM builds on these previous works by introducing memory modules and memory compression mechanisms, significantly reducing GPU consumption. Despite these advancements, most of these models are primarily focused on retaining more information from video. However, there is still relatively little research directly modeling the complex spatio-temporal dynamics of video, and improving a model's ability to capture these complex spatio-temporal representations remains a significant challenge.

2.2 Diffusion Models for Representation Learning

Diffusion model [Ho *et al.*, 2020] is a type of deep generative model that uses the final state of a Markov chain starting from a standard Gaussian distribution to approximate the distribution of natural images. Although diffusion models are mainly designed for generation tasks, the denoising process allows for the learning of both low- and high-level features from the input data [Fuest *et al.*, 2024].

Recent research has demonstrated the application of diffusion models as representation learners [Xiang *et al.*, 2023; Chen *et al.*, 2024], with improved models learning better representations [Xiang *et al.*, 2023]. Although the representational power of diffusion models is grounded in well-established theoretical foundations, leveraging off-the-shelf

diffusion models for non-generative tasks poses significant challenges [Yang and Wang, 2023]. In previous research, to fully leverage the representations extracted by diffusion models, these methods [Xiang *et al.*, 2023] typically require architectures similar to time-conditioned U-Net, a specialized structure ill-suited for long video understanding tasks. Due to the task mismatch between the teacher model and the student model, applying diffusion for knowledge distillation presents a noteworthy challenge. To address this issue, we employ a novel adaptive multi-level alignment strategy, offering a method for integrating high-quality representations. This allows us to leverage diffusion-based supervision to enhance the model's capabilities.

3 Method

To address the challenges of modeling fine-grained temporal and spatial relationships in the video, as shown in Figure 2, we propose Diff-LMM. This novel framework integrates diffusion-based supervision into video MLLMs. Unlike traditional methods that directly use CLIP and Q-Former for visual feature extraction, our approach employs diffusion encoders at intermediate layers. This allows diffusion features to enhance CLIP representations through feature alignment and knowledge distillation loss.

3.1 Visual Encoding

Our method, inspired by cognitive processes [Wu *et al.*, 2022], sequentially processes video frames to manage long-term visual information effectively. Given a video with N frames, we utilize a pre-trained visual encoder to extract features from each frame, forming a sequence $S = [s_1, s_2, \ldots, s_N]$, where $s_n \in R^{P\times C}$ represents the feature of the n-th frame, P is the number of image tokens, and C is the feature dimension per token. To enhance both temporal and spatial representation, we incorporate a temporal position encoding (PE) mechanism, integrating temporal dynamics with the frame-level features f_t:

$$f_t = s_n + \mathrm{PE}(t), \tag{1}$$

where $f_t \in R^{P\times C}$. The visual features are organized into a hierarchical memory bank using a dynamic compression strategy. This strategy reduces redundancy while retaining essential temporal features, ensuring rich information preservation and computational efficiency—key for long-term dynamic analysis.

Querying Transformer with Memory Banks. Previous studies [Zhang *et al.*, 2023; He *et al.*, 2024] have introduced the Querying Transformer (Q-Former) to capture temporal dynamics across frames and multimodal alignment between visual cues and queries in videos. The Q-Former models long time series using learnable query vectors $z \in R^{N\times C}$, where N is the number of queries and C is the feature dimension.

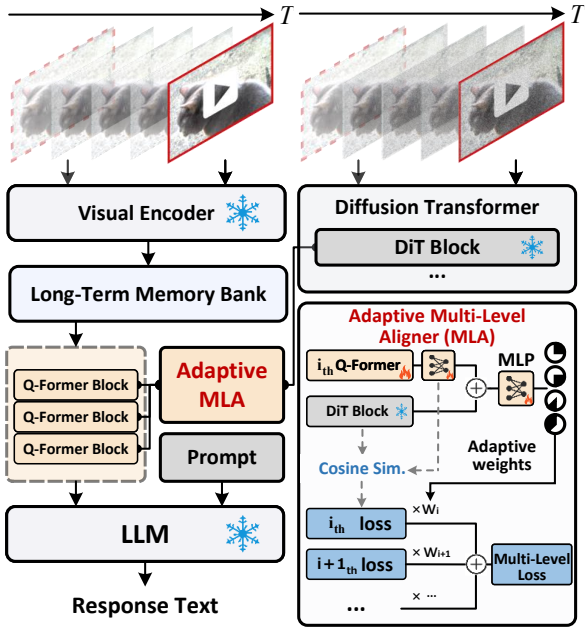

Figure 2: **Framework overview.** Diff-LMM processes video frames sequentially, using a visual encoder for feature extraction and the Q-Former to generate queries, which are stored in a long-term memory bank. The final output is obtained by decoding the Q-Former output from the last timestep using a pre-trained large model. To improve the model's ability to capture fine-grained spatio-temporal dynamics, we leverage a pre-trained diffusion model to extract frame-wise representations as supervisory signals. An adaptive multi-level alignment mechanism dynamically adjusts the loss weights based on semantic differences among Q-Former layers. During training, components with fixed parameters are marked with a snowflake icon, while tunable parts are indicated by a flame.

To improve the dynamic modeling of long-term temporal information, we adopt the long-term memory mechanism [He *et al.*, 2024], which combines both visual and query memory banks. The *visual memory bank* stores feature representations of all historical frames extracted by the frozen visual encoder, capturing long-range contextual information. At each time step t, this bank aggregates features and serves as the key-value input for the cross-attention mechanism in the Q-Former. The *query memory bank* captures temporal dynamics by dynamically accumulating query vectors learned by the Q-Former at each time step.

3.2 Learning from Diffusion Teacher

We contend that diffusion models possess a superior ability to capture spatial attributes compared to CLIP models, owing to their design and training objectives. While CLIP employs contrastive learning to align image and text embeddings by maximizing similarity between paired data, its emphasis on high-level semantic alignment may neglect finegrained spatial details [Monsefi *et al.*, 2024]. In contrast, diffusion models learn through iterative denoising of corrupted images, a process that necessitates precise reconstruction of spatial structures and textures.

In the context of multimodal large language models, while Q-Former efficiently enhances the original CLIP representa-

tion, it primarily aggregates global visual features via learnable query. This prevents the model from capturing fine-grained temporal and spatial relationships in videos. To address this issue, we introduce diffusion-based supervision for more detailed dynamic spatio-temporal modeling.

Pre-Trained Diffusion Models. Diffusion models, leveraging large-scale image data and visual generative priors, excel in extracting detailed and structural object cues [Wang *et al.*, 2023; Wang *et al.*, 2024a]. By incorporating feedback from these models, the fine-grained feature extraction capability of CLIP models is significantly enhanced [Wang *et al.*, 2024a]. These probabilistic models [Rombach *et al.*, 2022; Brooks *et al.*, 2024] learn the data distribution $p(x)$ and generate x from a random Gaussian variable, where x represents an image in the context of image diffusion models. To capture complex visual concepts, diffusion models reconstruct signals from noisy data x_τ at varying noise levels. The loss function for this process is defined as:

$$\mathcal{L}_{\text{diffusion}} = E_{x,\epsilon\in\mathcal{N}(0,1),\tau}\left[\|\epsilon - \epsilon_\theta(x_\tau, t)\|_2^2\right], \quad (2)$$

where ϵ represents the actual noise contaminating the clean data, and $\epsilon_\theta(x_\tau, t)$ indicates the noise predicted by the denoising model.

Noisy data x_τ is generated by adding noise from a Gaussian distribution $\mathcal{N}(0,1)$ to the clean data x_0, following the noise scheduler α_t [Ho *et al.*, 2020], as defined:

$$x_\tau = \sqrt{\alpha_\tau}x_0 + \sqrt{1-\alpha_\tau}\epsilon, \quad \epsilon \in \mathcal{N}(0,1). \quad (3)$$

Here, τ denotes the timestep in the diffusion process, with higher τ corresponding to increased noise.

Diffusion Supervision. While Q-Former excels in tasks such as Visual Question Answering (VQA), it struggles to represent detailed visual information due to CLIP's emphasis on low-frequency signals and global patterns [Park *et al.*, 2023]. To address this issue, we introduce supervisory signals from diffusion models, which have been shown to enhance CLIP's performance on fine-grained tasks [Wang *et al.*, 2024a]. Specifically, we utilize DiT to extract visual representations from each video frame, which serves as the predictive target for Q-Former hiddens. This alignment aims to improve Q-Former's capability to model intricate spatio-temporal dynamics effectively.

Let f be a pre-trained diffusion model, and consider a video frame x_t. The output of the diffusion model is $y_t = f(x_t) \in R^{H\times D}$ and the Q-Former output $h_t \in R^{P\times C}$, where H and D represent the number of hidden units and the embedding dimension, respectively. However, due to the architectural differences between the diffusion model and Q-Former, the representations extracted by the two models differ in both scale and dimension, making it challenging to align the outputs directly. Therefore, to reduce the difficulty of alignment training, we first apply average pooling to the output of the diffusion model, aiming to retain as much information as possible while reducing the dimensionality to improve alignment efficiency, resulting in the final alignment target $y^* \in \text{Pool}(f(x_t))$.

Furthermore, to address the differences in scale and dimension between the two models' outputs, we introduce a trainable projection head h_ϕ, which projects the Q-Former output h_t into the same dimensional space as the diffusion model's output, the resulting aligned feature is $h_\phi(h_t) \in R^D$. Finally, we align the projected Q-Former output $h_\phi(h_t)$ with the average pooled diffusion output y^* using a knowledge distillation loss function, which can be computed as follows:

$$\mathcal{L}_{kd} = -\frac{h_\phi(h_t)\cdot y^*}{\|h_\phi(h_t)\|\|y^*\|}, \quad (4)$$

where h_ϕ is parameterized using an MLP layer.

Using two linear projectors, we align the features from the Q-Former with those from the diffusion model despite the differences in their architectures. This effectively integrates valuable representations from the diffusion model as supervisory signals, which helps tackle the challenges associated with understanding long videos. However, while the above method resolves the alignment issue, ensuring training efficiency and improving model performance when aligning across multiple layers remains a challenge we face.

Adaptive Multi-Level Alignment. Due to potential semantic variations across the layers of Q-Former, excessive disparity between layers can lead to alignment failure. To optimize alignment and balance different semantic hierarchies, we propose an adaptive multi-level alignment module. This module performs dense alignment for each block and uses a learnable mechanism to assess alignment difficulty, improving learning efficiency and accuracy.

In particular, we utilize feature representations from a subset of Q-Former and integrate multi-level alignment losses through an adaptive weighting mechanism. Given a Q-Former with L layers, we select M levels for representation alignment. Let $\mathcal{U} = \{u_1, u_2, ..., u_M\}$ denote the indices of the selected Q-Former layers, where $M < L$ and $u_i \in \{1, 2, ..., L\}$, and $h_t^{(u_i)} \in R^{P\times C}$ represents the output of the u_i-th block. For each selected block, the alignment loss computed using Eq 4 is:

$$\mathcal{L}_{kd}^{(u_i)} = -\frac{h_{\phi_i}(h_t^{(u_i)})\cdot y^*}{\|h_{\phi_i}(h_t^{(u_i)})\|\|y^*\|}. \quad (5)$$

In the training process, a common approach involves multiplying each layer's alignment loss by a uniform weight and summing them to obtain the final alignment loss for parameter optimization. However, we observe that this method is suboptimal due to semantic differences in the representations extracted by different layers of the Q-Former. Assigning the same weight to all layers not only fails to achieve proper alignment but also impairs the Q-Former's ability to extract semantic features. To enable efficient and differentiated alignment across layers, we introduce a weight prediction network, g_ψ, which dynamically predicts the loss of weight for each layer. This prediction is based on the projection of the Q-Former output and the Diffusion alignment feature. Specifically, the alignment weights for different Q-Former blocks is computed as follows:

$$w_{\mathcal{U}} = \sigma(g_\psi(h_t^{\mathcal{U}} + y^*)), \quad (6)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function to normalize the weights. Then, the total representation alignment loss

can be computed as:

$$\mathcal{L}_{kd} = \sum_{i=1}^{M} w_{u_i} \mathcal{L}_{kd}^{(u_i)}. \tag{7}$$

In this way, our alignment mechanism dynamically adjusts the loss weight for each layer based on the different semantics extracted by each layer, achieving differential alignment. This provides an effective solution for balancing the preservation of the model's original performance while leveraging diffusion-based supervision.

3.3 Language Decoding

Finally, given the output features of the Q-Former and the text prompt (if possible), a frozen LLM is used to auto-regressively decode them into discrete text sequences. In summary, the training objective of our framework includes two main components: (1) a cross-entropy loss for video-to-language auto-regressive generation, and (2) a knowledge distillation loss for feature representation alignment. The cross-entropy loss is defined as:

$$\mathcal{L}_{ce} = -\frac{1}{S} \sum_{i=1}^{S} \log P(o_i | o_{<i}, V), \tag{8}$$

where o_i denotes the i-th token in the ground-truth sequence, and $P(o_i|o_{<i}, V)$ represents the model's prediction given the video context V and preceding tokens $o_{<i}$.

Similarly, to avoid damaging the model's performance due to the introduction of supervisory signals, it is crucial to balance both components. To achieve this, we introduce a hyperparameter λ in the final loss. The final loss is a weighted combination of both components:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{kd}, \tag{9}$$

where λ is a hyper-parameter that modulates the contribution of the representation alignment loss $\mathcal{L}_{kd}$, balancing their relative importance during training.

4 Experiments

4.1 Dataset

We evaluate our proposed method by assessing its performance in long video understanding, fine-grained video understanding, and video question-answering tasks.

Long Video Understanding. We conduct experiments on the LVU dataset [Wu and Krahenbuhl, 2021], which contains approximately 30K video clips from 3K movies, each with a duration ranging from 1 to 3 minutes. The dataset covers a diverse range of real-world scenarios, making it an extensive benchmark for long video recognition tasks. Our evaluation focuses on specific subtasks, such as director style prediction, and movie genre classification, fully leveraging the dataset's diversity and complexity to validate the effectiveness of the proposed method. Through these tasks, we further demonstrated the significant advantages of the proposed method in handling long video content, particularly in capturing temporal dependencies and contextual information. Additionally, by comparing with existing methods on this dataset, we revealed the limitations of current short-duration models when dealing with long video tasks.

Fine-Grained Video Understanding. To validate the capability of our proposed framework in modeling complex spatio-temporal dynamics, we conduct experiments focusing on more fine-grained subtasks in the LVU dataset [Wu and Krahenbuhl, 2021], i.e., relation recognition. As shown in Figure 5, the relation recognition task labels the relationships between two individuals as either husband and wife, friend, or other friendship. Even humans struggle to accurately distinguish these relationships based solely on a few video stills, as this requires inferring fine-grained features [Li *et al.*, 2024a] such as facial expressions and behaviors. To test whether our framework enhances the fine-grained video understanding ability of MLLMs, we focus on relationships between "husband and wife" and "friends", excluding "boyfriend and girlfriend" due to limited samples. This setup allows us to assess whether our model improves fine-grained spatio-temporal modeling.

Video Question Answering. To further compare with existing multimodal methods, we also extend the evaluation to the MSVD-QA [Xu *et al.*, 2017], a standard open-ended video question-answering dataset, consisting of short videos lasting 10-15 seconds. This dataset covers a variety of scenes and includes rich question-answering tasks, widely used to assess the multimodal understanding ability of video question answering systems. Through experiments on MSVD-QA, we not only validated the effectiveness of the proposed method in short-duration video understanding but also compared the performance of our Transformer-based architecture with that of current state-of-the-art methods in this task. The experimental results demonstrate that the proposed method also performs excellently in short-duration video question answering tasks, effectively combining visual and linguistic information, significantly improving the accuracy and robustness of video question answering.

4.2 Implementation Details

This study utilizes the Vicuna-7B model as the LLM. It is trained over 20 epochs with a learning rate of 1×10^{-4} and a batch size of 64. The AdamW optimizer is applied with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the hyperparameters, and a weight decay of 0.05. Input images are resized to 224×224 pixels. As for the diffusion model, we follow the REPA [Yu *et al.*, 2024] to use the eighth layer output of the diffusion model (DiT-XL-2-256×256) as the supervisory signal y^*. The input images of diffusion models is resized to 256×256 pixels. We assign a class label of 1000 (no class), and initialize the timestep to 0 (no noise) before inputting the images into the pre-trained, frozen DiT.

On the LVU dataset [Wu and Krahenbuhl, 2021], we set λ to 1, while on the MSVD-QA dataset [Xu *et al.*, 2017], λ is set to 0.0005. During the decoding phase, a beam search width of 5 is employed. The frame length is 100, while the memory bank length is set to 20. For Language-Video Understanding (LVU) tasks, the prompt format is "What is the $< task >$ of the movie?", where the task represents relationship, speaking style, scene, director, genre, writer, and release year. For evaluation, we choose the widely used top-1 accuracy as the primary metric.

Model	Content			Metadata				Avg
	Relationship	Speak	Scene	Director	Genre	Writer	Year	
Obj_T4mer [Wu and Krahenbuhl, 2021]	54.8	33.2	52.9	47.7	52.7	36.3	37.8	45.0
Performer [Choromanski *et al.*, 2020]	50.0	38.8	60.5	58.9	49.5	48.2	41.3	49.6
Orthoformer [Patrick *et al.*, 2021]	50.0	38.3	66.3	55.1	55.8	47.0	43.4	50.8
VideoBERT [Sun *et al.*, 2019]	52.8	37.9	54.9	47.3	51.9	38.5	36.1	45.6
VIS4mer [Islam and Bertasius, 2022]	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7
MA-LMM [He *et al.*, 2024]	58.2	44.8	**80.3**	**74.6**	61.0	70.4	**51.9**	63.0
Diff-LMM (Ours)	**69.2**	**45.2**	**80.3**	71.6	**63.2**	**75.0**	48.96	**64.8**

Table 1: Comparison with state-of-the-art methods on LVU. **Bold** and underline represent the best and second-best results.

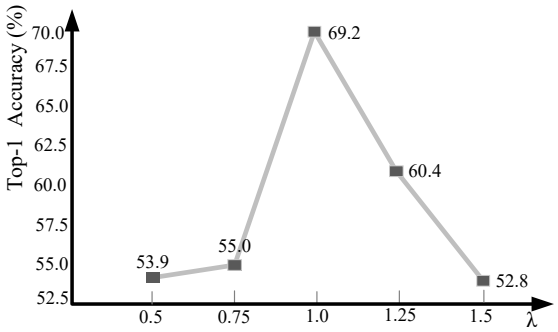

Figure 3: Ablation of hyperparameter λ on the relationship task of the LVU dataset.

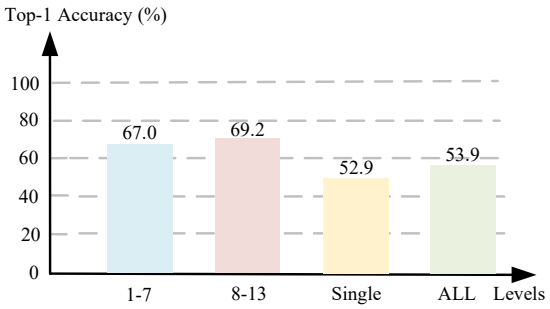

Figure 4: Analysis of the different alignment levels in Q-Former on the relationship task of the LVU dataset.

Model	Husband & Wife	Friend
MA-LMM [He *et al.*, 2024]	27.8	76.7
Diff-LMM (Ours)	**66.7**	**90.7**

Table 2: Comparison with Baselines on LVU [Wu and Krahenbuhl, 2021] fine-grained relationship subsets. Top-1 accuracy is reported.

Model	MSVD
GiT [Wang *et al.*, 2022]	56.8
mPLUG-2 [Xu *et al.*, 2023]	58.1
UMT-L [Li *et al.*, 2023b]	55.2
Video-LLaMA [Zhang *et al.*, 2023]	58.3
MA-LMM [He *et al.*, 2024]	60.6
Diff-LMM (Ours)	**60.8**

Table 3: Comparison with state-of-the-art methods on MSVD.

4.3 Comparision with State-of-the-arts

Long Video Understanding. As shown in Table 1, compared to previous methods (e.g., VideoBERT [Sun *et al.*, 2019], ViS4mer [Islam and Bertasius, 2022]) on the LVU benchmark [Wu and Krahenbuhl, 2021], our method achieves significant breakthroughs in both content understanding and metadata prediction tasks. This resulted in significant improvements in most tasks, enhancing the average top-1 accuracy by 1.8% compared to state-of-the-arts. While MA-LMM [He *et al.*, 2024] improves model performance on long video tasks by introducing advanced memory mechanisms, the CLIP model's tendency to prioritize global feature alignment [Park *et al.*, 2023] during training limits its performance on fine-grained tasks. In contrast, by incorporating diffusion supervision and feature alignment, our method enhances the model's ability to capture fine-grained spatio-temporal dynamics, enabling Diff-LMM to achieve state-of-the-art performance in fine-grained video understanding tasks.

Fine-Grained Video Understanding. In the LVU dataset [Wu and Krahenbuhl, 2021], relationship recognition requires identifying connections between characters in nearly 2.5-minute videos. This task is challenging due to its fine-grained nature, as it involves interpreting long-term spatial and temporal features [Li *et al.*, 2024a] like body movements and facial expressions. Distinguishing between spousal and friendship relationships can be particularly difficult, especially among heterosexual friends and spouses, leading to potential confusion without a subtle understanding of the video content. As demonstrated in Table 2, MA-LMM currently stands as a strong baseline method for long video understanding tasks. While it achieves an accuracy of 76.7% in recognizing friendships, it only manages to reach 27.8% accuracy in identifying spousal relationships, emphasizing the difficulty

of this particular task.

In contrast, leveraging diffusion-based supervision, our method achieves a Top-1 accuracy of 66.7% in recognizing spousal relationships, outperforming MA-LMM by about 38.9%. For friendship recognition, accuracy improves to 90.7%. As illustrated in Figure 5, our approach effectively distinguishes between heterosexual friendships and spousal relationships, which can be challenging for humans due to their behavioral similarities. This strong performance in fine-grained long video understanding supports the notion that diffusion-based supervision significantly enhances the model's ability to capture subtle spatio-temporal dynamics.

Video Question Answering. To validate the advantages of the proposed architecture, we compared it with existing multimodal video understanding methods on the widely used MSVD-QA dataset [Xu *et al.*, 2017], testing the robustness of spatio-temporal representations extracted through our method. Although Diff-LMM was not specifically designed for short video question-answering tasks, it is noteworthy that our model achieved state-of-the-art performance, as shown in Table 3. While the improvements on the MSVD-QA dataset are relatively modest, this indirectly highlights that fine-grained spatio-temporal modeling is more critical for long video understanding tasks than for short video tasks. Our approach significantly enhances this capability by incorporating diffusion-based supervision.

4.4 Ablation Studies

The Impact of λ. We analyze the impact of various λ values on the relationship task in the LVU datasets [Wu and Krahenbuhl, 2021], as shown in Figure 3. Setting λ to 1 yields the best performance for the adaptive multi-level alignment, allowing better alignment of representations from the diffusion model and overcoming the limitations of the original CLIP model, which mainly focuses on coarse, global alignment.

However, it is crucial to balance the degree of alignment between global and detailed features, as indicated in Figure 3. When λ is excessively high, the model becomes overly influenced by the diffusion representations. This causes the Q-Former to concentrate too much on detailed alignment in visual-text alignment, leading CLIP to overlook important global features and consequently reducing the model's accuracy. A similar issue arises when λ is set too low. Thus, we chose to set $\lambda = 1$ in our experiments to ensure optimal visual-text alignment for the CLIP model.

Alignment Level. We investigated the impact of applying alignment strategies to different layers of Q-Former. Our approach included aligning all layers, aligning a single level (specifically, the 13th layer), aligning the lower levels (layer 1-7), and aligning the higher levels (layer 8-13). Our findings indicate that aligning the higher layers during training yields the best results, as illustrated in Figure 4. Specifically, aligning the higher layers improves top-1 accuracy by nearly 15.3% compared to aligning all layers. The alignment of the lower layers produces results that are only slightly less effective than those of the higher layers. The reason for this is that when aligning all layers with the diffusion-extracted representations, the model ends up with similar hidden states

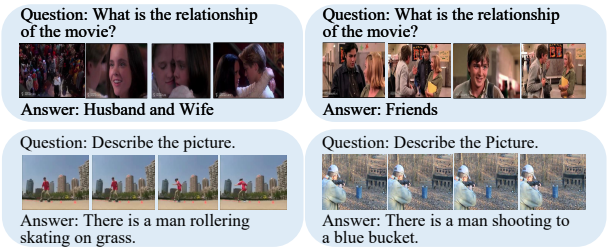

Figure 5: Visualization results of our method on long video recognition task on the LVU dataset.

w	**Adaptive**	**Average**
Top-1 Acc.	**69.2**	58.2

Table 4: Analysis of the adaptive loss mechanism's contribution to the relationship task of the LVU dataset.

across all layers of the Q-Former. This similarity hinders the model's ability to capture temporal dynamics present in the visual information of the video. Therefore, our method focuses on aligning only higher layers.

Adaptive Loss. To evaluate the effectiveness of the proposed adaptive loss mechanism, we conducted ablation experiments on the relationship task using the LVU dataset [Wu and Krahenbuhl, 2021]. The goal was to assess the necessity of the adaptive mechanism in balancing loss weights across different layers. For comparison, we implemented an alternative strategy where the weights were averaged, and the weight w_{u_i} for all layers remained constant. We evaluate both methods on the relationship task. As shown in Table 4, the adaptive loss mechanism outperforms the equal-weight strategy by 11%. This improvement is attributed to the variability in hidden states across layers, as aligning diffusion representations alone does not lead to better performance. Our method dynamically adjusts the layer weights, enabling more effective feature alignment.

5 Conclusion

In this paper, we propose a novel approach of introducing diffusion-based supervision for long video understanding. We investigate whether the representations generated by diffusion models can be effectively aligned with recent video MLLMs to address the limitations of models that neglect fine-grained information, which hinders their ability to capture detailed spatio-temporal dynamics. To mitigate this issue, we propose an adaptive multi-level alignment mechanism with diffusion supervision for robust and precise feature alignment. Experimental results show that pre-trained diffusers are powerful representation models, and our proposed alignment mechanism effectively improves the performance of MLLMs in complex, dynamic scenarios.

Contribution Statement

Jisheng Dang and Ligen Chen contributed equally to this work. Teng Wang is the corresponding author.

References

[Brooks *et al.*, 2024] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 2024.

[Chen *et al.*, 2024] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024.

[Choromanski *et al.*, 2020] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[Dang and Yang, 2021] Jisheng Dang and Jun Yang. Higcnn: Hierarchical interleaved group convolutional neural networks for point clouds analysis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2825–2829. IEEE, 2021.

[Dang and Yang, 2022] Jisheng Dang and Jun Yang. Lhphgcnn: Lightweight hierarchical parallel heterogeneous group convolutional neural networks for point cloud scene prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):18903–18915, 2022.

[Dang *et al.*, 2023a] Jisheng Dang, Huicheng Zheng, Jinming Lai, Xu Yan, and Yulan Guo. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *IEEE Transactions on Image Processing*, 32:3924–3938, 2023.

[Dang *et al.*, 2023b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, and Yulan Guo. Unified spatio-temporal dynamic routing for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(5):4512–4526, 2023.

[Dang *et al.*, 2024a] Jisheng Dang, Huicheng Zheng, Bimei Wang, Longguang Wang, and Yulan Guo. Temporo-spatial parallel sparse memory networks for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Dang *et al.*, 2024b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, and Yulan Guo. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *IEEE Transactions on Image Processing*, 2024.

[Dang *et al.*, 2024c] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, Qingyong Hu, and Yulan Guo. Adaptive sparse memory networks for efficient and robust video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Fuest *et al.*, 2024] Michael Fuest, Pingchuan Ma, Ming Gui, Johannes S Fischer, Vincent Tao Hu, and Bjorn Ommer. Diffusion models and representation learning: A survey. *arXiv preprint arXiv:2407.00783*, 2024.

[He *et al.*, 2024] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Islam and Bertasius, 2022] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[Li *et al.*, 2023a] 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*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2023b] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023.

[Li *et al.*, 2024a] Lin Li, Jun Xiao, Guikun Chen, Jian Shao, Yueting Zhuang, and Long Chen. Zero-shot visual relation detection via composite visual cues from large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Li *et al.*, 2024b] Xiangxian Li, Yuze Zheng, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. Cross-modal learning using privileged information for long-tailed image classification. *Computational Visual Media*, 10(5):981–992, 2024.

[Ma *et al.*, 2024] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. Triple sequence learning for cross-domain recommendation. *ACM Transactions on Information Systems*, 42(4):1–29, 2024.

[Maaz *et al.*, 2023] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[Meng *et al.*, 2024] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Meng *et al.*, 2025] Lei Meng, Xiangxian Li, Xiaoshuo Yan, Haokai Ma, Zhuang Qi, Wei Wu, and Xiangxu Meng. Causal inference over visual-semantic-aligned graph for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19449–19457, 2025.

[Monsefi *et al.*, 2024] Amin Karimi Monsefi, Kishore Prakash Sailaja, Ali Alilooee, Ser-Nam Lim, and Rajiv Ramnath. Detailclip: Detail-oriented clip for fine-grained tasks. *arXiv preprint arXiv:2409.06809*, 2024.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Park *et al.*, 2023] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? *arXiv preprint arXiv:2305.00729*, 2023.

[Patrick *et al.*, 2021] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

[Ren *et al.*, 2024] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Song *et al.*, 2024] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[Sun *et al.*, 2019] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[Wang *et al.*, 2022] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[Wang *et al.*, 2023] Yule Wang, Zijing Wu, Chengrui Li, and Anqi Wu. Extraction and recovery of spatio-temporal structure in latent dynamics alignment with diffusion models. *Advances in Neural Information Processing Systems*, 36:38988–39005, 2023.

[Wang *et al.*, 2024a] Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps clip see better. *arXiv preprint arXiv:2407.20171*, 2024.

[Wang *et al.*, 2024b] Yuqing Wang, Lei Meng, Haokai Ma, Yuqing Wang, Haibei Huang, and Xiangxu Meng. Modeling event-level causal representation for video classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3936–3944, 2024.

[Wu and Krahenbuhl, 2021] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021.

[Wu *et al.*, 2022] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.

[Xiang *et al.*, 2023] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023.

[Xu *et al.*, 2017] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[Xu *et al.*, 2023] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023.

[Yang and Wang, 2023] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023.

[Yu *et al.*, 2024] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[Zong *et al.*, 2024] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.