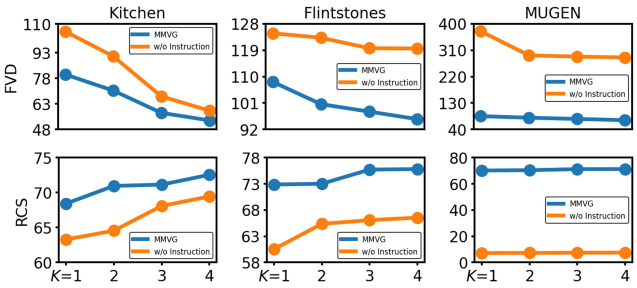


Figure 1. Different **numbers of past frame** for TVPrediction (**MMVG** *vs.* **w/o Instruction**).

Method	T-VQ	VidSwin	Kitchen FVD↓	Kitchen RCS↑	Flintstones FVD↓	Flintstones RCS↑	MUGEN FVD↓	MUGEN RCS↑
TATS [4]	-	-	87.2	66.3	115.9	70.6	90.1	67.9
MMVG	✗	✗	**81.5**	**68.1**	**110.1**	**72.4**	**86.3**	**69.6**
	✓	✗	80.7	68.3	108.4	72.6	85.7	70.0
MMVG	✗	✓	80.8	68.0	108.6	72.3	86.0	69.9
	✓	✓	**80.2**	**68.4**	**108.2**	**72.9**	**84.8**	**70.2**

Table 1. **Ablation study** of MMVG with temporal-aware VQGAN (T-VQ) and VideoSwin decoder (VidSwin) for TVPrediction.

A. More Past Frames for TVPrediction

As illustrated in Fig. 1, we explore the effect of different past frames (K) for TVP. With more past frames, we see a noticeable improvement for MMVG w/o Instruction (*e.g.*, FVD decreases from 400 to 250 on MUGEN, and RCS increases from 63 to 68 on Kitchen). However, it is still far behind the one that has language guidance. Even only the first frame with the text outperforms using 4 past frames on Flintstones (*e.g.*, a lower 110 FVD and a higher 73 RCS). Furthermore, MUGEN performs a series of actions and requires a longer temporal coherence; 4 past frames are insufficient to tell the expected outcome (*e.g.*, the poor 7 RCS). The above results demonstrate the cruciality of instruction. On the other hand, for humans, supplying language is easier than drawing more video frames. Our TVC provides a practical setting that leads to effective video completion performance as well as human efficiency.

B. Ablation Study

We conduct an ablation study to investigate each component effect in MMVG, including temporal-aware VQGAN (T-VQ) and VideoSwin decoder (VidSwin). T-VQ makes the reconstructed video from discrete tokens more smooth, and VidSwin considers latent temporal during the video decoding. If without T-VQ and VidSwin, MMVG will share a similar model architecture to TATS [4] but contain the proposed masking strategy that learns video completion from arbitrary frames. In Table 1, the performance gain mainly comes from the masking strategy (*e.g.*, a lower 81.5 FVD on Kitchen and a higher 69.6 RCS on MUGEN), which validates the core idea of the mask-then-recover learning. Both T-VQ and VidSwin benefit the temporal coherence of video modeling, leading to an FVD decrease with a slight increase in RCS. In addition, combining all of them can bring a comprehensive improvement to MMVG.

C. Detailed Analysis

All experiments are conducted on the **Kitchen** dataset for **TVPrediction**. We have a detailed comparison among VQGAN [8], TA-VQ [4], and T-VQ. As shown below, our T-VQ outperforms the others on both frame reconstruction (over real video) and further TVP.

VQ Model	Reconstruction MSE↓	Reconstruction FVD↓	TVPrediction FVD↓	TVPrediction RCS↑
VQGAN [8]	0.01582	36.72	82.3	66.4
TA-VQ [4]	0.00926	20.05	80.8	68.0
T-VQ (ours)	**0.00868**	**14.83**	**80.2**	**68.4**

The Dec^{Q} predicts the next frame based on the previous ones in an autoregressive manner. This enables a smooth transition and models the temporal dependencies, which is crucial for generative video modeling [4, 7, 18]. We also consider using parallel decoding, but it performs far below the autoregressive way.

Decoding	FVD↓	RCS↑
Parallel	113.5	64.6
Autoregressive	**80.2**	**68.4**

For training, we have a high masking ratio p, with no more than 4 frames. Each frame is 64 tokens, so the length is no longer than 64*4+5([SPAN])+77(text)=340, which is efficient for current sequential modeling. During inference, there could be only two frames (head and tail) for the infilling task. We also compare the entire generation process (enc+dec+VQ) to VideoDiff [5]. MMVG shows a better time/GPU efficiency for single and parallel inference.

Model	Time (sec) BS=1	Time (sec) 4	GPU (MB) BS=1	GPU (MB) 4
VideoDiff [5]	20.6	45.0	22419	37400
MMVG	**16.9**	**20.1**	**19814**	**32018**

We also try applying the same cube embedding and optimize it via MSE loss. The result indicates that MMVG still surpasses VideoMAE [13]. Moreover, the cube embedding cannot present clear and detailed pixels, which is unsuitable for video generation.

Model	Output	FVD↓	RCS↑
VideoMAE [13]	Cube	328.9	47.6
MMVG^{U}	Cube	**272.6**	**50.7**
MMVG^{U}	VQ	**105.6**	**63.3**

For a fair comparison to TATS [4], we adopt the same 1024 codebook size and a 24-layer transformer. We consider the common setting, 8192 codes (as DALL-E) or 12 layers (as BERT-base). More VQ codes do not affect the

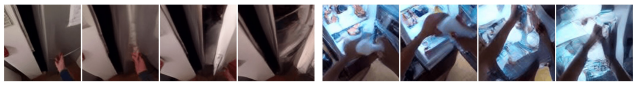
Method	FT.	Kitchen R@1	Kitchen R@5	Flintstones R@1	Flintstones R@5	MUGEN R@1	MUGEN R@5
CLIP [9]	✗	2.0	7.4	23.0	45.0	0.2	1.6
	Mean	11.4	39.2	73.2	97.0	11.4	31.2
	Temporal	**33.6**	**79.8**	**93.4**	**100**	**47.2**	**84.4**

Table 2. Results of **instruction-to-video retrieval** by CLIP with different fine-tunings (FT.). We sample 1K pairs for this study.

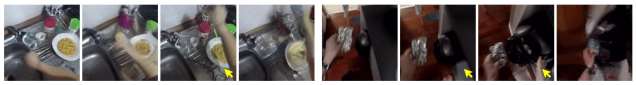
quality. In contrast, to imitate the diverse activity motions, MMVG requires a larger model capability. Without teacher-forcing, the model cannot be trained effectively since it associates incorrect inputs with the corresponding outputs.

Codebook	#Layer	FVD↓	RCS↑
1024	12	99.3	65.2
1024	24	**80.2**	**68.4**
8192	24	79.9	68.3

For TVC, it is challenging to generate new objects that are not presented in the visual cues or the instruction. As the failure cases, though MMVG can make the motion of "*open the fridge*", the items inside the fridge remain blurry. This highlights the need for human common sense to achieve realistic video generation.



We also demonstrate video completion from intermediate, where ↖ means the provided frames.



D. Fine-tune CLIP as Evaluator

The CLIP model [9] has shown promising results by its strong text-visual alignment. GDOVIA [17] adopts CLIP and first proposes relative CLIP similarity (RCS) to evaluate text-guided visual generation. Since the video scene is in a specific domain and may differ from CLIP, we further fine-tune CLIP on each TVC dataset for a more precise alignment as our evaluator. We consider two fine-tuned settings: *Mean* and *Temporal*. *Mean* applies a mean pooling layer over the visual features of all frames as the video features. On the other hand, *Temporal* incorporates LSTM [6] to acquire the temporal video features over frame features. Table 2 presents the instruction-to-retrieval results with different fine-tuned settings. A higher recall represents a better alignment between instructions and videos. If directly using CLIP for RCS, it results in poor performance and is insufficient for our evaluation. By considering the latent temporal within video features, *Temporal* leads to an overall advance and brings reliable alignment. We then finalize our evaluator as CLIP with the *Temporal* fine-tuning.

Method	Pre-training	UCF-101 IS↑	UCF-101 FVD↓
CogVideo [7]	5.4M	50.5	626
Make-A-Video [11]	20M	82.6	81
TATS [4]	✗	71.6	341
MMVG	✗	**73.7**	**328**

Table 3. Results of **text-to-video generation** on UCF-101. We follow CogVideo [7] to treat class labels as the input text. We gray out methods that use significantly more pre-training data.

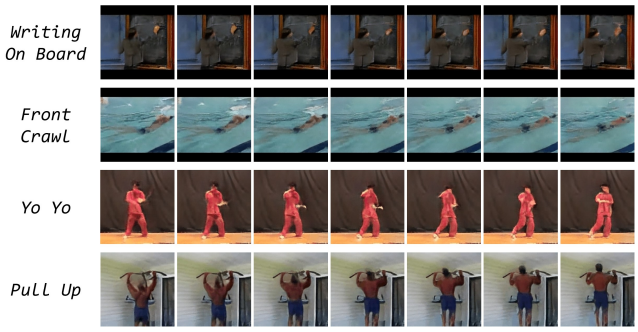


Figure 2. Qualitative examples of **text-to-video** on UCF-101.

E. Text-to-Video Generation on UCF-101

We follow CogVideo [7] to treat class labels as the input text for text-to-video generation on UCF-101 [12]. The results are shown in Table 3. Our MMVG, without additional training data, can surpass large-scale pre-trained CogVideo. The higher 73.7 IS shows that the generated results are more diverse [10]. And the lower 328 FVD also supports its better temporal coherence to ground-truth videos. When comparing MMVG to TATS, our masking strategy indicates the effectiveness that learning from completion can improve text-to-video. The qualitative examples are illustrated in Fig. 2.

F. Inferior Qualitative Results by VideoDiff

We show that diffusion methods cannot generate as high-quality video as the used visual-token transformers (*e.g.*, higher FVDs by VideoDiff [5] and MCVD [15]). We further illustrate the qualitative examples by VideoDiff in Fig. 3. As more challenging natural videos, we can see the blurring scenes on Kitchen. The motions are also unclear to tell what is actually doing. For Flintstones, it can produce characters but is difficult to present temporal dynamics, where the videos look almost static. Since it attempts to generate video frames from the 3D auto-encoder, VideoDiff cannot handle temporal coherence well. We still find obvious inconsistent results on MUGEN, even with the autoregressive video extension (*e.g.*, the agent disappears with the platform being different lengths in the first case, or the ladder wrongly shows up in the third row.).

Figure 3. Qualitative examples of **unconditional video generation** on Kitchen, Flintstones, and MUGEN by VideoDiff [5].

G. Human Evaluation

As illustrated in Fig. 4, we investigate the quality of generated results from the human aspect via Amazon Mechanical Turk. MTurkers rank the correlation of the TVC result

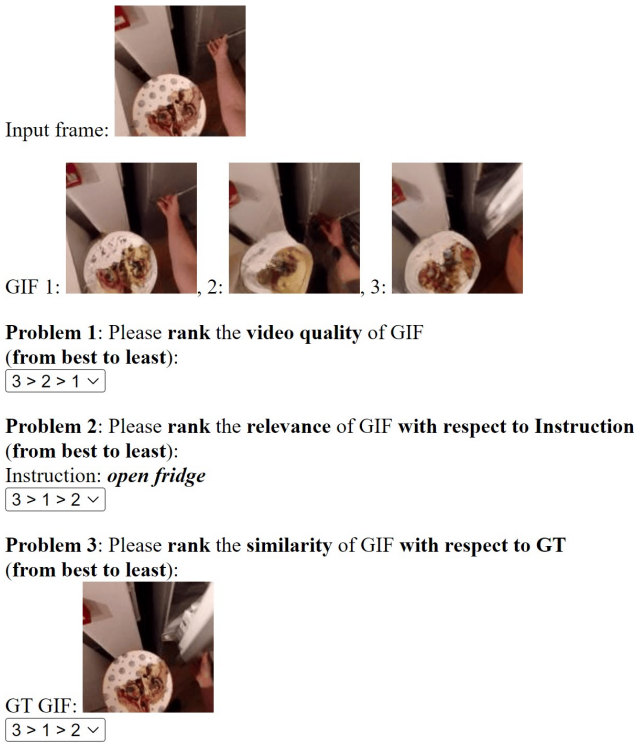


Figure 4. Screenshot of the ranking tasks for human evaluation.

concerning video quality, instruction relevance, or ground-truth similarity. Each MTurker rewards $4.0 for a question group and takes a mean of 7 minutes.

H. Social Impact and Ethics Discussion

TVC brings out a general video completion that can generate a video from frames at arbitrary time points and control via natural language. Although our work benefits creative visual applications, there may be a "*fake as real*" doubt for those produced videos. To mitigate this issue, we follow techniques in image forensics [3, 16] and train a binary classifier [14] to detect video authenticity. The accuracy on Kitchen, Flintstones, and MUGEN are all >99%, which prevents them from counterfeiting. For guided instructions, hate speech detection [1, 2] can be adopted to filter out potential malicious texts to avoid controversial results.

References

[1] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep Learning Models for Multilingual Hate Speech Detection. In *ECML-PKDD*, 2020. 3

[2] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Annual Meetings of the Association for Computational Linguistics Workshop (ACLW)*, 2021. 3

[3] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging

Frequency Analysis for Deep Fake Image Recognition. In *ICML*, 2020. 3

[4] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[5] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models. In *arXiv:2204.03458*, 2022. 1, 2, 3

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. In *Neural Computation*, 1997. 2

[7] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In *arXiv:2205.15868*, 2022. 1, 2

[8] Björn Ommer Patrick Esser, Robin Rombach. Taming Transformers for High-Resolution Image Synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016. 2

[11] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *arXiv:2209.14792*, 2022. 2

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. In *arXiv:1212.0402*, 2012. 2

[13] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[14] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[15] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[16] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images are Surprisingly Easy to Spot...for Now. In *CVPR*, 2020. 3

[17] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions. In *arXiv:2104.14806*, 2021. 2

[18] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video Generation using VQ-VAE and Transformers. In *arXiv:2104.10157*, 2021. 1