VACE: All-in-One Video Creation and Editing

Zeyinzi Jiang* Zhen Han* Chaojie Mao*† Jingfeng Zhang Yulin Pan Yu Liu
Tongyi Lab, Alibaba Group

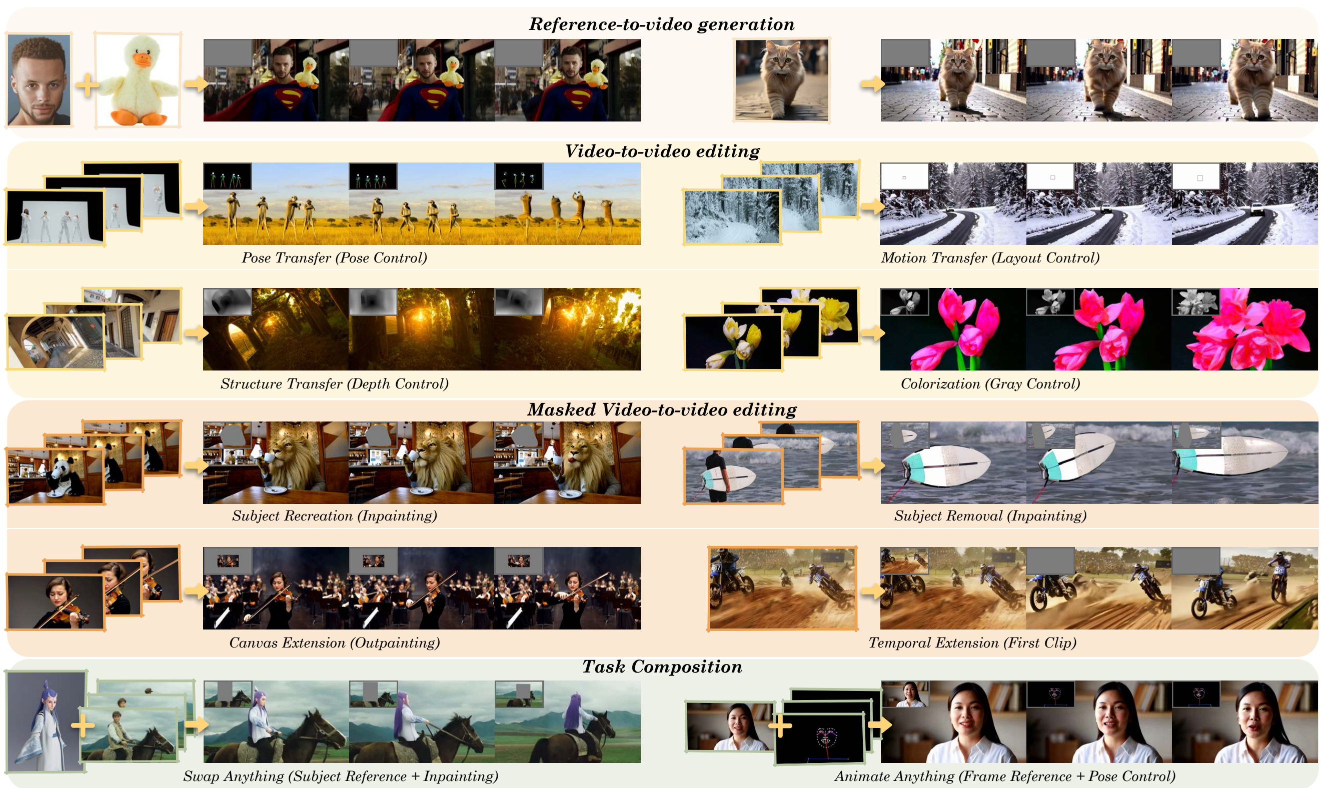


Figure 1. **Comprehensive capabilities of `VACE`.** We present the outstanding generation results based on Wan2.1 (left) and LTX-Video (right). For each task, the original input image or video (left), the context video (top left corner), and the generated frames are illustrated.

Abstract

*Diffusion Transformer has demonstrated powerful capability and scalability in generating high-quality images and videos. Further pursuing the unification of generation and editing tasks has yielded significant progress in the domain of image content creation. However, due to the intrinsic demands for consistency across both temporal and spatial dynamics, achieving a unified approach for video synthesis remains challenging. We introduce **VACE**, which enables users to perform **V**ideo tasks within an **A**ll-in-one framework for **C**reation and **E**diting. These tasks include reference-to-video generation, video-to-video editing, and masked video-to-video editing. Specifically, we effectively integrate the requirements of various tasks by organizing video task inputs, such as editing, reference, and masking, into a unified interface referred to as the Video Condition Unit (VCU). Furthermore, by utilizing a Context Adapter structure, we inject different task concepts into the model using formalized representations of temporal and spatial dimensions, allowing it to handle arbitrary video synthesis tasks flexibly. Extensive experiments demonstrate that the unified model of VACE achieves performance on par with task-specific models across various subtasks. Simultaneously, it enables diverse applications through versatile task combinations. Project page:* `https://ali-vilab.github.io/VACE-Page/`.

*Equal Contribution. †Project lead.

1. Introduction

In recent years, the domain of visual generation tasks has witnessed remarkable advancements, driven in particular by the rapid evolution of diffusion models [24, 25, 48, 53, 54, 56, 57]. Beyond the early foundational pre-trained models for text-to-image [7, 16, 33] or text-to-video [9, 22, 64] generation in the field, there has been a proliferation of downstream tasks and applications, such as repainting [3, 82], editing [4, 42, 68, 70, 75], controllable generation [30, 76], frame reference generation [20, 73], and ID-referenced video synthesis [11, 35, 47, 74]. This array of developments highlights the dynamic and complex nature of the visual generation field. To enhance task flexibility and reduce the overhead associated with deploying multiple models, researchers have begun to focus on constructing unified model architectures [12, 63] (*e.g.*, ACE [23, 41] and OmniGen [71]), aiming to integrate different tasks into a single image model, facilitating the creation of various application workflows while maintaining simplicity in usage. In the field of video, due to the collaborative transformations in both temporal and spatial dimensions, leveraging a unified model can present endless possibilities for video creation. However, leveraging diverse input modalities and ensuring spatiotemporal consistency are still challenging for unified video generation and editing.

We propose **`VACE`**, an all-in-one model for video creation and editing that performs tasks including reference-to-video generation, video-to-video editing, masked video-to-video editing, and free composition of these tasks, as illustrated in Fig. 1. The aggregation of various capabilities reduces the costs of service deployment and user interaction, and by combining the capabilities of different tasks within a single model, it addresses challenges faced by existing video generation models, such as controllable generation of long videos, multi-condition and reference-based generation, and continuous video editing, thereby empowering users with greater creativity. To achieve this, we utilize the current mainstream Diffusion Transformers (DiTs) structure as the foundational video framework and pre-trained text-to-video generation models [22, 64], which provide better basic capabilities and scalability for handling long video sequences. Specifically, `VACE` takes into account the needs of different tasks during its construction and designs a unified interface, dubbed the Video Condition Unit (VCU), which integrates multiple modalities such as images or videos for editing, references, and masks. Furthermore, to differentiate the visual modality information in editing and reference tasks, we introduce the concept decoupling strategy, enabling the model to understand what aspects need to be retained and what should be modified. Meanwhile, by employing a pluggable Context Adapter structure, concepts from different tasks (*e.g.*, the areas or ranges of editing or reference) are injected into the model through collaborative spatiotemporal representation, enabling it to possess the capability of adaptive processing for unified tasks.

Due to the lack of existing multi-task benchmarks in video synthesis, we construct a dataset of 480 evaluation samples containing 12 different tasks, while evaluating the performance of the `VACE` unified model by comparing it with existing specialized models. Experimental results demonstrate that our framework exhibits sufficient competitiveness in both quantitative and qualitative analyses. To the best of our knowledge, we are the first all-in-one model based on the video DiT architecture that concurrently supports such a wide range of tasks. Notably, this innovative framework allows for the compositional expansion of base tasks, enabling the construction of scenarios such as long video re-rendering, which provides a versatile and efficient solution for video synthesis, opening new possibilities for user-side video content creation and editing.

2. Related Work

Visual Generation and Editing. With the rapid development of image [2, 7, 16, 18, 58, 59] and video [22, 32, 73, 77] generation models, they are being used to create high-quality visual content and are widely applied in fields such as advertising, film special effects, game development, and animation production [13, 43–45, 55]. Meanwhile, to meet the diverse needs of visual media production and to enhance efficiency and quality, precise generation and editing methods have emerged. Models are required to perform generative creation based on multimodal inputs, such as depth, structure, pose, scene, and characters. According to the purposes of the input conditions, we can categorize them into two types: editing of the input and concept-guided re-creation. A significant portion of the work, such as ControlNet [76], ControlVideo, Composer [26], VideoComposer [68], and SCEdit [30], focuses on single-condition editing and multi-condition composite editing based on temporal and spatial alignment conditions. Additionally, some works that focus on interactive local editing scenarios, such as DragGAN [46] and MagicBrush [75]. Methods that guide generation based on semantic information from the input, such as Cone [38], Cone2 [39], InstantID [67], and PuLID [21], can achieve conceptual understanding of the input and inject it into the model for creative purposes.

Task-unified Visual Generative Model. As the complexity and diversity of user creations increase, relying solely on a single model or a complicated chain of multiple models can no longer provide a convenient and efficient path for implementing creative ideas. In image generation, a unified generation and editing framework has begun to emerge, allowing for more flexible creative approaches. Methods such as UltraEdit [81] and SEED-Data-Edit [19] provide general-purpose editing datasets, while techniques like In-

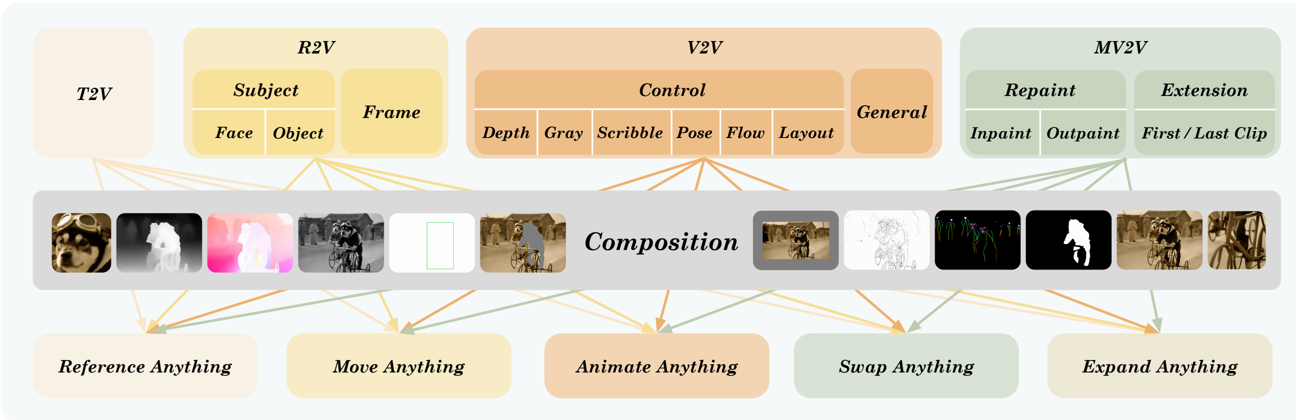


Figure 2. **Task categories covered by VACE.** Four basic tasks can be combined to create a vast number of possibilities.

structPix2Pix [4], MagicBrush [61], and CosXL [60] offer general instruction-based editing features. Additionally, methods like UniControl [50] and UNIC-Adapter [15] have unified controllable generation. Further advancements have led to the development of ACE [23, 41], OmniGen [71], OmniControl [63], and UniReal [12], which expand the scope of tasks by providing flexible controllable generation, local editing, and reference-guided generation. In the video domain, due to the increased difficulty of generation, approaches often manifest as single-task single-model frameworks, offering capabilities for editing or reference generation, as seen in Video-P2P [37], MagicEdit [34], MotionCtrl [69], Magic Mirror [80], and Phantom [35]. VACE aims to fill the gap for a unified model within the video domain, providing possibilities for complex creative scenarios.

3. Method

VACE is designed as a multimodal-to-video generation model, where text, image, video, and mask are integrated into a unified conditioning input. To cover as many video generation and editing tasks as possible, we conduct in-depth research into existing tasks, then divide them into 4 categories according to their individual requirements of multimodal inputs. Without losing generality, we specifically design a novel multimodal input format for each category under a Video Condition Unit (VCU) paradigm. Finally, we restructure the DiT model for VCU inputs, making it a versatile model for a wide range of video tasks.

3.1. Multimodal Inputs and Video Tasks.

Despite existing video tasks varying in complex user inputs and ambitious creative goals, we found that most of their inputs can be fully represented in 4 modalities: text, image, video, and mask. Overall, as illustrated in Fig. 2, we group these video tasks into 5 categories based on their requirements of these four multimodal inputs.

- **Text-to-Video Generation (T2V)** is a basic video creation task, and text is the only input.
- **Reference-to-Video Generation (R2V)** requires additional images as reference inputs, making sure that specified contents, such as subjects of faces, animals, and other objects, or video frames, appear in the generated video.
- **Video-to-Video Editing (V2V)** makes an entire change to a provided video, such as colorization, stylization, controllable generation, *etc*. We use video control types whose control signals can be represented and stored as RGB videos, including depth, grayscale, pose, scribble, optical flow, and layout; however, the method itself is not limited to these.
- **Masked Video-to-Video Editing (MV2V)** makes changes to an input video only within the provided 3D regions of interest (3D ROI), seamlessly blending in with the other unchanged regions, such as inpainting, outpainting, video extension, *etc*. We use an extra spatiotemporal mask to represent the 3D ROI.
- **Task Composition** includes all the compositional possibilities of the 4 types of video tasks above.

3.2. Video Condition Unit

We propose an input paradigm, Video Condition Unit (VCU), to unify diverse input conditions into textual input, frame sequence, and mask sequence. A VCU can be denoted as

$$V = [T; F; M], \tag{1}$$

where T is a text prompt, while F and M are sequences of context video frames $\{u_1, u_2, ..., u_n\}$ and masks $\{m_1, m_2, ..., m_n\}$ respectively. Here, u is in RGB space, normalized to $[-1, 1]$ and m is binary, in which "1"s and "0"s symbolize where to edit or not. F and M are aligned both in spatial size $h \times w$ and temporal size n. In T2V, no context frame or mask is required. To keep generality, we assign a default value $0_{h\times w}$ to each u denoting empty input, and set every m to $1_{h\times w}$ meaning that all these pixels

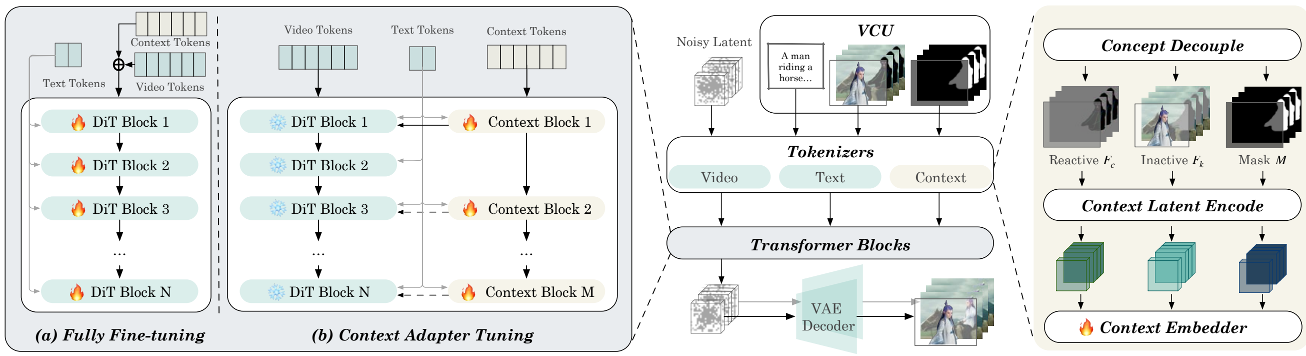


Figure 3. **Overview of VACE Framework.** Frames and masks are tokenized through Concept Decoupling, Context Latent Encode, and Context Embedder. To achieve training with VCU as input, we employ two strategies: (a) Fully Fine-tuning and (b) Context Adapter Tuning. The latter converges faster and supports pluggable features.

Table 1. The formal representation of frames (Fs) and masks (Ms) under the four basic tasks. Frames and masks are aligned spatially and temporally.

Tasks	Frames (Fs) & Masks (Ms)
T2V	$F = \{0_{h\times w}\} \times n$ $M = \{1_{h\times w}\} \times n$
R2V	$F = \{r_1, r_2, ..., r_l\} + \{0_{h\times w}\} \times n$ $M = \{0_{h\times w}\} \times l + \{1_{h\times w}\} \times n$
V2V	$F = \{u_1, u_2, ..., u_n\}$ $M = \{1_{h\times w}\} \times n$
MV2V	$F = \{u_1, u_2, ..., u_n\}$ $M = \{m_1, m_2, ..., m_n\}$

are about to be re-generated. For R2V, additional reference frames r_i are inserted in front of the default frame sequence, while all-zero masks $0_{h\times w}$ are inserted in front of the mask sequence. These all-zero masks mean that the corresponding frames should be kept unchanged. In V2V, the context frame sequence is the input video frames, and the context mask is a sequence of $1_{h\times w}$. For MV2V, both context video and mask are required. The formal mathematical representations are shown in Tab. 1.

VCU can also support task composition. For example, the context frames of reference-inpainting task are $\{r_1, r_2, ..., r_l, u_1, u_2, ..., u_n\}$ and the context masks are $\{0_{h\times w}\} \times l + \{m_1, m_2, ..., m_n\}$. In this case, users can modify l objects in the video and regenerate based on the provided reference images. For another example, users only have a scribble image and want to generate a video beginning with the contents described by this scribble image, which is a scribble-based video extension task. The context frames are $\{u\} + \{0_{h\times w}\} \times (n-1)$ and the context masks are $\{1_{h\times w}\} \times n$. In this way, we can achieve multi-condition and reference control generation for long videos.

3.3. Architecture

We restructure the DiT model for VACE, as shown in Fig. 3, aiming to support multimodal VCU inputs. Since there is an existing pipeline for text tokenization, we only consider the tokenization of context frames and masks. After being tokenized, the context tokens are combined with noisy video tokens and used to fine-tune the DiT model. Different from that, we also propose a Context Adapter Tuning strategy, which allows context tokens to pass through Context Blocks and be added back to the original DiT Blocks.

3.3.1. Context Tokenization

Concept Decoupling. Two different visual concepts of natural video and control signals, like depth, pose, are encoded in F simultaneously. We believe that explicitly separating these data of different modalities and distributions is essential for model convergence. The concept decoupling is based on masks and yields two frame sequences identical in shape: $F_c = F \times M$ and $F_k = F \times (1 - M)$, where F_c is called the reactive frames contain all the pixels to be changed, while all the pixels to be kept are stored in F_k, named inactive frames. Specifically, the reference images and the unchanged part of V2V and MV2V go to F_k, while control signals and those pixels about to change, such as gray pixels, are collected to F_c.

Context Latent Encoding. A typical DiT processes noisy video latents $X \in \mathbb{R}^{n' \times h' \times w' \times d}$, where n', h', and w' are the temporal and spatial shapes of the latent space. Similar to X, F_c, F_k, and M need to be encoded into a high-dimensional feature space to ensure the property of significant spatiotemporal correlations. Therefore, we reorganize them together with X into a hierarchical and spatiotemporal aligned visual features. F_c, F_k are processed by video VAE and mapped into the same latent space of X, maintaining their spatiotemporal consistency. To avoid any unexpected mishmash of images and videos, reference images are separately encoded by VAE encoder and concatenated back along the temporal dimension, while the corresponding parts need to be removed during decoding. M is directly reshaped and interpolated. After that, F_c, F_k, and M are all mapping into latent spaces and are spatiotemporal aligned with X in the shape of $n' \times h' \times w'$.

Context Embedder. We extend the embedder layer by concatenating F_c, F_k, and M in the channel dimension and tokenizing them into context tokens, which we refer to as the Context Embedder. The corresponding weights to tokenize F_c and F_k are directly copied from the original video embedder, and the weights to tokenize M are initialized to zeros.

3.3.2. Fully Fine-Tuning and Context Adapter Tuning

To achieve training with VCU as input, a simple methodology is to fully fine-tune the whole DiT model, as shown in Fig. 3a. Context tokens are added together with noisy tokens X, and all the parameters in DiT and the newly introduced Context Embedder will be updated during training. To avoid fully fine-tuning and achieve faster convergence, as well as to establish a pluggable feature with the foundation model, we also propose another methodology processing the context token in a Res-Tuning [29] manner, as shown in Fig. 3b. Particularly, we select and copy several Transformer Blocks from the original DiT, forming a distributed and cascade-type Context Blocks. The original DiT processes video tokens and text tokens, while the newly added Transformer Blocks process context tokens and text tokens. The output of each Context Block is inserted back into the DiT blocks as an additive signal to assist the main branch in performing generation and editing tasks. In this manner, the parameters of DiT are frozen. Only the Context Embedder and Context Blocks are trainable.

4. Datasets

4.1. Data Construction

To obtain an all-in-one model, the diversity and complexity of the required data construction also increase. Existing common text-to-video and image-to-video tasks only require constructing pairs of text and video. However, for the tasks in `VACE`, the modalities need to be further expanded to include target videos, source videos, local masks, reference, and more. To efficiently and rapidly acquire data for various tasks, it is imperative to maintain video quality while also conducting instance-level analysis and understanding of the video data.

To achieve this, we first analyze the video data itself by performing shot slicing and preliminarily filtering out data based on resolution, aesthetic score, and motion amplitude. Next, we label the first frame of the video using RAM [78] and combine it with Grounding DINO [36] for detection, utilizing the localization results to perform secondary filtering on videos with target areas that are either too small or too large. Furthermore, we employ the propagate operation of SAM2 [52] for video segmentation to obtain instance-level information across the video. Leveraging the results of video segmentation, we filter instances in the temporal dimension by calculating the effective frame ratio based on the mask area threshold.

In the actual training process, the construction for different tasks also needs to be tailored according to the characteristics of each task: **1)** For some controllable video generation tasks, we pre-extract depth [51], scribble [6], pose [5, 72], and optical flow [65] from the filtered videos. For gray and layout tasks, we create data on the fly. **2)** For repainting tasks, random instances from the videos can be masked for inpainting, while the inverse of the mask enables the construction of outpainting data. Augmentation of the masks [62] allows for unconditional inpainting. **3)** In the case of extension tasks, we extract key frames such as the first frame, last frame, frames from both ends, random frames, and segments from both ends to support a wider variety of extension types. **4)** For reference tasks, we can extract several face or object instances from the videos and apply offline or online augmentation operations to create paired data. Notably, we randomly combine all the previously mentioned tasks for training to accommodate a broader range of model application scenarios. Additionally, for all operations involving masks, we perform arbitrary augmentation to satisfy various granular local generation requirements.

4.2. VACE-Benchmark

Significant progress has been made in the field of video generation. However, a scientific and thorough evaluation of the performance of these models remains an urgent issue that needs to be addressed. VBench [27] and VBench++ [28] have established a precise evaluation framework for text-to-video and image-to-video tasks through an extensive assessment suite and dimensional design. Nevertheless, as the video generation ecosystem continues to evolve, more derivative tasks have begun to emerge, such as video reference generation and video editing, for which a comprehensive benchmark is still lacking. To address this gap, we propose `VACE-Benchmark` to evaluate various downstream tasks related to video in a systematic manner.

Starting from the data sources, we recognize that real videos and generated videos may exhibit different performance characteristics during evaluation. Thus, we collected a total of 240 high-quality videos categorized by their sources, encompassing various data types, including text-to-video, inpainting, outpainting, extension, grayscale, depth, scribble, pose, optical flow, layout, reference face, and reference object tasks, with an average of 20 samples for each task. The input modalities include videos, masks, and references, and we also provide the original videos

Table 2. **Quantitative evaluations on `VACE-Benchmark`.** We compare the automated score metrics of the unified `VACE` based on LTX-Video and the proprietary model on the dimensions of video quality and video consistency, as well as results of human user studies.

Type	Method	Video Quality & Video Consistency									User Study			
		Aesthetic Quality	Background Consistency	Dynamic Degree	Imaging Quality	Motion Smoothness	Overall Consistency	Subject Consistency	Temporal Flickering	Normalized Average	Prompt Following	Temporal Consistency	Video Quality	Average
I2V	I2VGenXL [77]	55.20%	92.87%	**60.00%**	63.31%	97.43%	23.78%	89.58%	95.67%	71.54%	2.65	1.60	2.34	2.20
	CogVideoX-I2V [73]	**57.78%**	94.80%	40.00%	**68.23%**	98.69%	24.38%	**93.84%**	97.84%	73.66%	**3.30**	2.28	**3.19**	2.92
	LTX-Video [22]	56.12%	94.57%	35.00%	62.72%	**99.27%**	24.92%	92.83%	**98.41%**	72.89%	2.95	2.28	2.28	2.50
	`VACE` (Ours)	57.53%	**95.32%**	45.00%	68.03%	99.08%	**25.13%**	93.61%	97.80%	**74.38%**	3.20	**4.00**	2.54	**3.24**
Inpaint	ProPainter [82]	44.70%	95.64%	**50.00%**	**61.57%**	99.01%	18.48%	92.99%	**98.47%**	70.15%	2.35	**4.00**	**2.99**	**3.11**
	`VACE` (Ours)	**51.30%**	**96.30%**	**50.00%**	60.39%	**99.12%**	**21.12%**	**94.59%**	98.21%	**72.05%**	**2.40**	**4.00**	2.60	3.00
Outpaint	Follow-Your-Canvas [8]	53.30%	95.99%	5.00%	**69.53%**	98.08%	**25.90%**	**95.38%**	97.20%	71.54%	3.05	2.00	1.63	2.23
	M3DDM [17]	53.34%	95.87%	**30.00%**	65.07%	**99.22%**	25.43%	93.65%	**98.85%**	73.16%	3.70	3.88	2.28	3.29
	`VACE` (Ours)	**57.04%**	**96.55%**	**30.00%**	69.49%	99.20%	25.36%	94.47%	98.47%	**74.25%**	**3.90**	**3.92**	**3.58**	**3.80**
Depth	Control-A-Video [10]	50.62%	91.71%	**70.00%**	**67.76%**	97.58%	24.48%	88.10%	96.58%	72.35%	2.70	2.28	1.54	2.17
	VideoComposer [68]	50.03%	94.18%	**70.00%**	59.44%	96.23%	24.95%	89.79%	94.38%	70.74%	2.60	2.44	2.17	2.40
	ControlVideo [79]	**63.30%**	95.02%	10.00%	65.13%	96.49%	24.20%	92.29%	95.42%	70.07%	2.55	2.50	1.82	2.29
	`VACE` (Ours)	56.72%	**96.12%**	60.00%	66.41%	**98.84%**	**25.27%**	**94.09%**	**97.27%**	**74.99%**	**3.10**	**3.92**	**2.66**	**3.23**
Pose	Text2Video-Zero [31]	57.63%	87.67%	**100.00%**	**70.74%**	79.65%	23.94%	84.82%	76.57%	59.69%	2.15	2.00	1.88	2.01
	ControlVideo [79]	**65.37%**	94.56%	25.00%	65.28%	97.32%	25.19%	92.76%	**96.82%**	72.45%	2.15	1.80	2.03	1.99
	Follow-Your-Pose [40]	48.79%	86.80%	**100.00%**	67.41%	90.12%	26.10%	80.18%	88.02%	66.43%	2.00	2.60	1.58	2.06
	`VACE` (Ours)	60.17%	**94.92%**	75.00%	64.71%	**98.63%**	**26.44%**	**94.82%**	96.60%	**76.13%**	**2.95**	**3.96**	**2.63**	**3.18**
Flow	FLATTEN [14]	**56.23%**	95.80%	70.00%	61.65%	97.86%	**26.23%**	93.94%	96.17%	74.42%	**3.50**	2.40	**3.19**	3.03
	`VACE` (Ours)	55.76%	**96.07%**	**75.00%**	**65.37%**	**98.98%**	25.89%	**94.63%**	**96.93%**	**75.90%**	2.90	**3.75**	2.60	**3.08**
R2V	Kling1.6 [1]	62.13%	96.04%	**85.00%**	69.27%	99.38%	**27.82%**	93.79%	97.79%	**78.81%**	**4.22**	**4.10**	3.80	**4.04**
	Pika2.2 [49]	62.48%	96.79%	65.00%	69.87%	99.37%	26.02%	95.93%	98.90%	77.87%	4.00	3.85	**3.87**	3.91
	Vidu2.0 [66]	**64.30%**	96.85%	35.00%	67.03%	**99.66%**	26.53%	96.73%	**99.41%**	76.47%	3.90	3.85	3.77	3.84
	`VACE` (Ours)	63.25%	**98.03%**	30.00%	**72.29%**	99.51%	25.85%	**98.54%**	99.15%	76.76%	3.47	3.42	3.30	3.40

to enable further processing by developers based on the specific characteristics of each task. Regarding the data prompts, we supply the original captions of the videos for quantitative assessment, as well as rewritten prompts tailored to the specific tasks to evaluate the models' creativity.

5. Experiments

5.1. Experimental Setup

Implementation Details. `VACE` is trained based on Diffusion Transformers for text-to-video generation at various scales. It utilizes LTX-Video-2B [22] for faster generation, while Wan-T2V-14B [64] is used specifically for higher-quality outputs, supporting resolutions of up to 720p. The training employs a phased approach. Initially, we focus on foundational tasks such as inpainting and extension, which are considered modal complementary to the pre-trained text-to-video models. This includes the incorporation of masks and the learning of contextual generation in both spatial and temporal dimensions. Next, from a task expansion perspective, we progressively transition from single input reference frames to multiple input reference frames and from single tasks to composite tasks. Finally, we fine-tune the model's quality using higher-quality data and longer sequences. The input for model training accommodates arbitrary resolutions, dynamic durations, and variable frame rates to support diverse input needs of users.

Baselines. Our goal is to achieve the unification of video creation and editing tasks, and currently, there is no comparable all-in-one video generation model available, which leads us to focus our evaluation on comparing our general model with proprietary task-specific models. Moreover, due to the numerous tasks involved and the lack of open-sourced methods for many of them, we conduct our comparisons on models that are available either offline or online. Specifically for the tasks, we compare the following: **1)** For the I2V task, we examine I2VGenXL [77], CogVideoX-I2V [73], and LTX-Video-I2V [22]; **2)** In the repainting task, we compare the ProPainter [82] for removal inpainting, while Follow-Your-Canvas [8] and M3DDM [17] are compared for outpainting; **3)** For controllable task, we use Control-A-Video [10], VideoComposer [68], and ControlVideo [79] under depth conditions, and compare Text2Video-Zero [31], ControlVideo [79], and Follow-Your-Pose [40] under pose conditions, as well as FLAT-

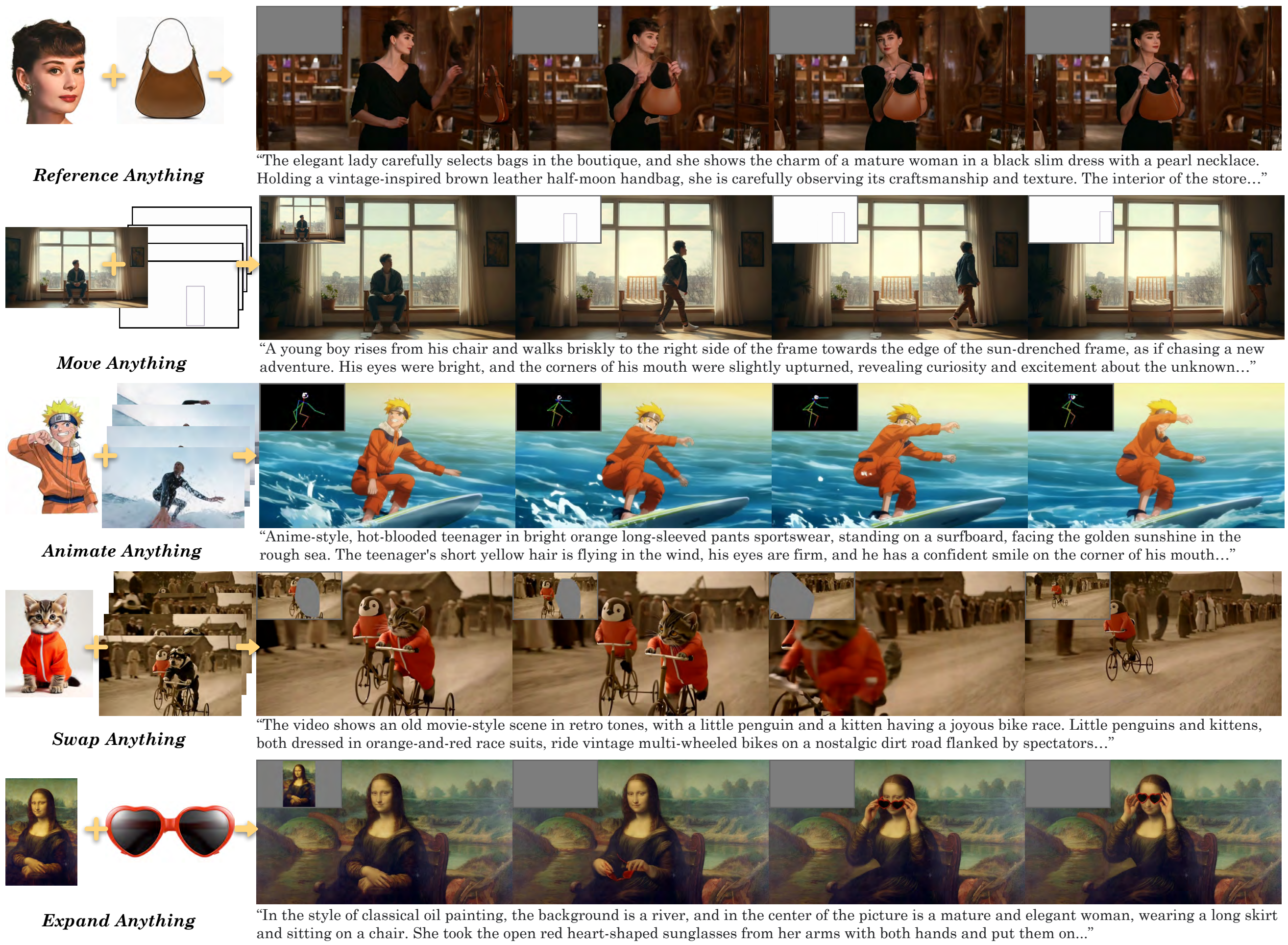


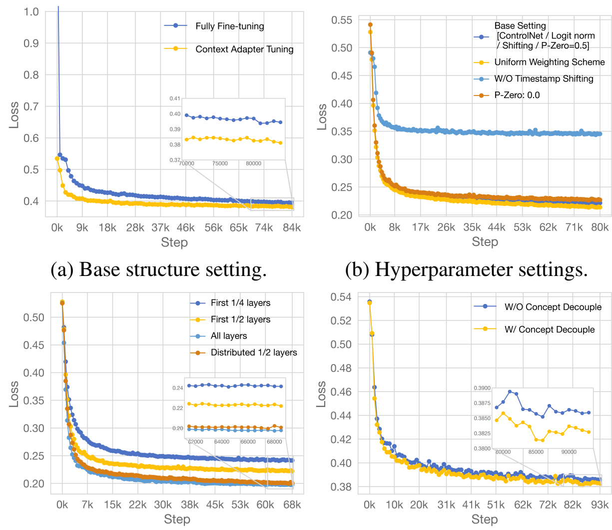
Figure 4. **Visualization results of compositional tasks.** `VACE` creatively enables reference-, move-, animate-, swap-, and expand-anything.

TEN [14] under optical flow conditions; **4)** In reference generation, given the absence of open-sourced models, we compare commercial products Kling1.6 [1], Pika2.2 [49], and Vidu2.0 [66].

Evaluation. To comprehensively evaluate the performance of various tasks, we employ the `VACE-Benchmark` for assessment. Specifically, we divide the evaluation into automatic scoring and a user study for manual assessment. For the automatic scoring, we utilize select metrics from VBench [27] to assess video quality and video consistency, including eight indicators: aesthetic quality, background consistency, dynamic degree, imaging quality, motion smoothness, overall consistency, subject consistency, and temporal flickering. For the manual assessment, we utilize the mean opinion score (MOS) as our evaluation metric, focusing on three aspects: prompt following, temporal consistency, and video quality. In practice, we anonymize the generated data and randomly distribute it to different participants for scoring on a scale from 1 to 5.

5.2. Main Results

Quantitative Evaluation. We compare `VACE` comprehensive model based on LTX-Video with task proprietary approaches on `VACE-Benchmark`. For certain tasks, we follow existing methods; for example, although we support generating based on any frame, we conduct comparisons using the first-frame reference approach from current open-source methods to ensure fairness. From Tab. 2, we can see that for the tasks of I2V, inpainting, outpainting, depth, pose, and optical flow, our method demonstrates better performance than other open-source methods across eight indicators of video quality and video consistency, with normalized average metrics showing superior results. Some competing methods can only generate at a resolution of 256, have very short generation durations, and exhibit instability in temporal coherence, resulting in poorer performance on automatic metric calculations. For the R2V task, there is still a certain gap in metrics compared to commercial models for a small-scale model that aims for fast generation,



(a) Base structure setting. (b) Hyperparameter settings.

(c) Context adapter configurations. (d) Concept decouple setting.

Figure 5. **Ablation Studies** of the `VACE` regarding structures, hyperparameters, and module configurations.

while being comparable to the metrics of Vidu 2.0. According to the results of human user studies, our method consistently performs better in evaluation metrics across multiple tasks, aligning well with user preferences.

Qualitative Results. In Fig. 1, we present the results of the `VACE` single model across various tasks. It is evident that the model achieves a high level of performance in video quality and temporal consistency. Furthermore, in composition tasks shown in Fig. 4, our model showcases impressive abilities, effectively integrating different modalities and tasks to produce results that cannot be generated by existing single or multiple models, thereby demonstrating its strong potential in the fields of video generation and editing. For example, in the “Move Anything” case, by providing a single input image and a movement trajectory, we are able to precisely move the characters in the scene with a specified direction while maintaining coherence and narrative consistency.

5.3. Ablation Studies

To better understand the impact of different independent modules on a unified video generation framework, we conducted a series of systematic comparative experiments based on the LTX-Video model to achieve a better model structure and configuration. To accurately assess the different experimental settings, we sample 250 data points for each task as a validation set and calculate the training loss, reflecting the model’s training progress through the mean curve changes of different tasks.

Base Structure. Text-guided image or video generation models only take noise as inference input. When extended to our unified input paradigm, VCU, we can conduct training using fully fine-tuning or by incorporating additional parameter fine-tuning. Specifically, as shown in Fig. 5a, we compare the concatenation of different inputs along the channel dimension and modify the input dimensions of the patchify projection layer to achieve the loading and fully fine-tuning of the pre-trained model. Additionally, we introduce some extra training parameters in the form of Res-Tuning [29], which serializes VCU in a bypass branch and injects information into the main branch. The results indicate that both methods yielded similar effects; however, since the additional parameter fine-tuning converges faster, we base our subsequent experiments on this approach. As shown in Fig. 5b, we further conduct hyperparameter experiments based on this structure, focusing on aspects such as weighting schemes, timestamp shifting, and p-zero.

Context Adapter. Since the number of context blocks will significantly affect the model size and inference time consumption, we attempt to find an optimal number and distribution of context blocks. We begin with selecting continuous blocks at the input side and make comparisons between the first 1/4 blocks, 1/2 blocks, and all blocks. Inspired by the Res-Tuning [29] method, we also experiment with evenly distributing the injection blocks instead of selecting a continuous block series. As shown in Fig. 5c, we can see that when using the same number of blocks, the distributed arrangement of blocks outperforms the continuous arrangement in shallow blocks. Furthermore, a greater number of blocks generally yields better results, but due to the limited improvement in effectiveness and the constraints of training resources, we adopt a partially distributed arrangement of blocks.

Concept Decouple. During training, we introduce a Concept Decouple processing module to further disassemble the visual units, clarifying what content the model needs to learn to modify or retain. As shown in Fig. 5d, using this module results in a more significant reduction in loss.

6. Conclusion

This paper introduces `VACE`, an all-in-one video generation and editing framework. It unifies the diverse and complex multimodal inputs required for various video tasks, bridging the gap between specialized models for each individual task. This enables most video AI creation tasks to be completed with a single inference of a single model. While broadly covering various video tasks, `VACE` also supports flexible and free combinations of these tasks, greatly expanding the application scenarios of video generation models and meeting a wide range of user creative needs. The `VACE` framework paves the way for the development of unified visual generative models with multimodal inputs and represents a significant milestone in the field of visual generation.

Acknowledgments. We would like to express our sincere appreciation for the contributions of many colleagues for their insightful discussions, valuable suggestions, and constructive feedback, including: Yuwei Wang, Haiming Zhao, Chenwei Xie, and Sheng Yao for their data contributions, and Shiwei Zhang, Tao Fang, and Xiang Wang for their discussions and suggestions.

References

[1] KLING AI. KLING AI, https://klingai.com/, 2025. 6, 7, 13

[2] Runway AI. Stable Diffusion v1.5 Model Card, https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022. 2

[3] Runway AI. Stable Diffusion Inpainting Model Card, https://huggingface.co/runwayml/stable-diffusion-inpainting, 2022. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning To Follow Image Editing Instructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18392–18402, 2023. 2, 3

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186, 2021. 5

[6] Caroline Chan, Frédo Durand, and Phillip Isola. Learning To Generate Line Drawings That Convey Geometry and Semantics. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7915–7925, 2022. 5

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2

[8] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-Your-Canvas: Higher-Resolution Video Outpainting with Extensive Content Generation. In *Assoc. Adv. Artif. Intell.*, 2025. 6, 13

[9] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, Ting-Che Lin, Shilong Zhang, Fu Li, Chuan Li, Xing Wang, Yanghua Peng, Peize Sun, Ping Luo, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Goku: Flow Based Video Generative Foundation Models. *arXiv preprint arXiv:2502.04896*, 2025. 2

[10] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-A-Video: Controllable Text-to-Video Diffusion Models with Motion Prior and Reward Feedback Learning. *arXiv preprint arXiv:2305.13840*, 2023. 6, 13

[11] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot Object-level Image Customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[12] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang Zhao. UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics. *arXiv preprint arXiv:2412.07774*, 2024. 2, 3

[13] Alibaba Cloud. Tongyi Wanxiang, https://tongyi.aliyun.com/wanxiang, 2023. 2

[14] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: Optical FLow-guided ATTENtion for consistent text-to-video editing. In *Int. Conf. Learn. Represent.*, 2024. 6, 7, 13

[15] Lunhao Duan, Shanshan Zhao, Wenjun Yan, Yinglun Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Mingming Gong, and Gui-Song Xia. UNIC-Adapter: Unified Image-instruction Adapter with Multi-modal Transformer for Image Generation. *arXiv preprint arXiv:2412.18928*, 2024. 3

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Int. Conf. Mach. Learn.*, 2024. 2

[17] Fanda Fan, Chaoxu Guo, Litong Gong, Biao Wang, Tiezheng Ge, Yuning Jiang, Chunjie Luo, and Jianfeng Zhan. Hierarchical Masked 3D Diffusion Model for Video Outpainting. In *ACM Int. Conf. Multimedia*, pages 7890–7900, 2023. 6, 13

[18] FLUX. FLUX, https://blackforestlabs.ai/, 2024. 2

[19] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. SEED-Data-Edit Technical Report: A Hybrid Dataset for Instructional Image Editing. *arXiv preprint arXiv:2405.04007*, 2024. 2

[20] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, Haibin Huang, and Chongyang Ma. I2V-Adapter: A General Image-to-Video Adapter for Diffusion Models. In *ACM SIGGRAPH*, pages 1–12, 2024. 2

[21] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. PuLID: Pure and Lightning ID Customization via Contrastive Alignment. In *Adv. Neural Inform. Process. Syst.*, 2024. 2

[22] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*, 2025. 2, 6, 13

[23] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. ACE: All-round Creator and Editor Following Instructions via Diffusion Transformer. In *Int. Conf. Learn. Represent.*, 2025. 2, 3

[24] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Adv. Neural Inform. Process. Syst.* Curran Associates, Inc., 2020. 2

[26] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and Controllable Image Synthesis with Composable Conditions. In *Int. Conf. Mach. Learn.*, 2023. 2

[27] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21807–21818, 2024. 5, 7

[28] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. *arXiv preprint arXiv:2411.13503*, 2024. 5

[29] Zeyinzi Jiang, Chaojie Mao, Ziyuan Huang, Ao Ma, Yiliang Lv, Yujun Shen, Deli Zhao, and Jingren Zhou. Res-Tuning: A Flexible and Efficient Tuning Paradigm via Unbinding Tuner from Backbone. In *Adv. Neural Inform. Process. Syst.*, 2023. 5, 8

[30] Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, and Jingfeng Zhang. SCEdit: Efficient and Controllable Image Diffusion Generation via Skip Connection Editing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8995–9004, 2024. 2

[31] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In *Int. Conf. Comput. Vis.*, pages 15954–15964, 2023. 6, 13

[32] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[33] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *arXiv preprint arXiv:2405.08748*, 2024. 2

[34] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. MagicEdit: High-Fidelity and Temporally Coherent Video Editing. *arXiv preprint arXiv:2308.14749*, 2023. 3

[35] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 2, 3

[36] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*, 2023. 5

[37] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-P2P: Video Editing with Cross-attention Control. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8599–8608, 2024. 3

[38] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept Neurons in Diffusion Models for Customized Generation. In *Int. Conf. Mach. Learn.*, 2023. 2

[39] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable Image Synthesis with Multiple Subjects. In *Adv. Neural Inform. Process. Syst.*, 2023. 2

[40] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Ying Shan, Xiu Li, and Qifeng Chen. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. In *Assoc. Adv. Artif. Intell.*, 2024. 6, 13

[41] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. ACE++: Instruction-Based Image Creation and Editing via Context-Aware Content Filling. *arXiv preprint arXiv:2501.02487*, 2025. 2, 3

[42] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *Int. Conf. Learn. Represent.*, 2021. 2

[43] Midjourney. Midjourney, https://www.midjourney.com, 2023. 2

[44] MiniMax. Hailuo AI Video, https://hailuoai.com/video, 2024.

[45] OpenAI. DALL·E 3, https://openai.com/dall-e-3, 2023. 2

[46] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. In *ACM SIGGRAPH*, 2023. 2

[47] Yulin Pan, Chaojie Mao, Zeyinzi Jiang, Zhen Han, and Jingfeng Zhang. Locate, Assign, Refine: Taming Customized Image Inpainting with Text-Subject Guidance. *arXiv preprint arXiv:2403.19534*, 2024. 2

[48] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Int. Conf. Comput. Vis.*, pages 4195–4305, 2023. 2

[49] PiKa. PiKa, https://pika.art/, 2025. 6, 7, 13

[50] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming

Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. In *Adv. Neural Inform. Process. Syst.*, 2023. 3

[51] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1623–1637, 2022. 5

[52] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *Int. Conf. Learn. Represent.*, 2025. 5

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 2

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Med. Image Comput. Computer-Assisted Interv.*, 2015. 2

[55] Runway. Gen-3, `https://app.runwayml.com/video-tools`, 2025. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *Int. Conf. Learn. Represent.*, 2021. 2

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *Int. Conf. Learn. Represent.*, 2021. 2

[58] StabilityAI. Stable Diffusion v2-1 Model Card, `https://huggingface.co/stabilityai/stable-diffusion-2-1`, 2022. 2

[59] StabilityAI. Stable Diffusion XL Model Card, `https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0`, 2022. 2

[60] StabilityAI. CosXL Model Card, `https://huggingface.co/stabilityai/cosxl`, 2024. 3

[61] Ya Sheng Sun, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, and Hideki Koike. ImageBrush: Learning Visual In-Context Instructions for Exemplar-Based Image Manipulation. In *Adv. Neural Inform. Process. Syst.*, 2023. 3

[62] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-Robust Large Mask Inpainting With Fourier Convolutions. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 2149–2159, 2022. 5

[63] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and Universal Control for Diffusion Transformer. *arXiv preprint arXiv:2411.15098*, 2024. 2, 3

[64] Wan Team. Wan: Open and advanced large-scale video generative models. 2025. 2, 6, 13

[65] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Eur. Conf. Comput. Vis.*, pages 402–419, 2020. 5

[66] Vidu. Vidu, `https://www.vidu.cn/`, 2025. 6, 7, 13

[67] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot Identity-Preserving Generation in Seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[68] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional Video Synthesis with Motion Controllability. In *Adv. Neural Inform. Process. Syst.*, 2023. 2, 6, 13

[69] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. In *ACM SIGGRAPH*, pages 1–11, 2024. 3

[70] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, Yingya Zhang, and Hongming Shan. DreamVideo-2: Zero-Shot Subject-Driven Video Customization with Precise Motion Control. *arXiv preprint arXiv:2410.13830*, 2024. 2

[71] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified Image Generation. *arXiv preprint arXiv:2409.11340*, 2024. 2, 3

[72] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective Whole-body Pose Estimation with Two-stages Distillation. In *Int. Conf. Comput. Vis.*, pages 4210–4220, 2023. 5

[73] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *Int. Conf. Learn. Represent.*, 2025. 2, 6, 13

[74] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-Preserving Text-to-Video Generation by Frequency Decomposition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025. 2

[75] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. In *Adv. Neural Inform. Process. Syst.*, 2023. 2

[76] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, 2023. 2

[77] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. *arXiv preprint arXiv:2311.04145*, 2023. 2, 6, 13

[78] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo,

Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize Anything: A Strong Image Tagging Model. *arXiv preprint arXiv:2306.03514*, 2023. 5

[79] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-free Controllable Text-to-Video Generation. In *Int. Conf. Learn. Represent.*, 2024. 6, 13

[80] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magic Mirror: ID-Preserved Video Generation in Video Diffusion Transformers. *arXiv preprint arXiv:2411.13503*, 2025. 3

[81] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. *arXiv preprint arXiv:2407.05282v1*, 2024. 2

[82] Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. ProPainter: Improving Propagation and Transformer for Video Inpainting. In *Int. Conf. Comput. Vis.*, pages 10477–10486, 2023. 2, 6, 13