MiniCausal-T2V: Towards Ultra-Low Latency and Memory-Efficient Causal Video Generation on Edge Devices

Bowen Long and Min-ho Kang *

Posted Date: 6 February 2026

doi: 10.20944/preprints202602.0468.v1

Keywords: text-to-video; Edge AI; diffusion model; low latency; causal transformer

 doi:10.20944/preprints202602.0468.v1

Article

MiniCausal-T2V: Towards Ultra-Low Latency and Memory-Efficient Causal Video Generation on Edge Devices

Bowen Long and Min-ho Kang

Jeonju University
* Correspondence: 2017112306@student.dongguk.edu

Abstract

The proliferation of Text-to-Video (T2V) generation technologies has opened new avenues for content creation, yet deploying these advanced models on resource-constrained edge devices remains a significant challenge due to their inherent complexity and high computational demands. This paper introduces MiniCausal-T2V (MCT-Video), an innovative, end-to-end optimized causal latent video diffusion model meticulously engineered for ultra-low latency and memory-efficient T2V generation on edge platforms, particularly Qualcomm Hexagon NPUs. MCT-Video distinguishes itself through a suite of synergistic innovations: a Lightweight Causal Transformer Backbone designed from scratch for intrinsic efficiency and causality, an Adaptive Sparse Temporal Attention mechanism for dynamic temporal computation reduction, Quantization-Aware Fine-tuning for robust precision deployment, a Unified Multi-objective Distillation strategy to holistically transfer knowledge, and Extreme Step Flow-Matching Inference for rapid generation. Extensive experimental evaluations demonstrate that MCT-Video not only achieves superior video quality across comprehensive VBench metrics and human perception but also sets new benchmarks for efficiency, achieving unprecedented end-to-end inference latency and a minimal memory footprint on Hexagon NPUs, substantially outperforming existing edge-optimized solutions. This work represents a significant step towards enabling high-quality, real-time T2V capabilities directly on portable devices.

Keywords: text-to-video; Edge AI; diffusion model; low latency; causal transformer

1. Introduction

The rapid evolution of Text-to-Video (T2V) generation has revolutionized digital content creation, enabling the synthesis of high-quality, high-resolution video sequences directly from textual prompts. This breakthrough holds immense potential across various domains, including automated content production, immersive AR/VR experiences, personalized media, and intelligent assistance systems, which can significantly impact small and medium-sized enterprises (SMEs) by optimizing budget allocation and growth strategies [1–3]. As digital content creation expands, the importance of content authenticity and protection also rises, necessitating advancements in areas like versatile image watermarking for tamper localization and copyright protection [4,5], and explainable forgery detection using multi-modal large language models [6].

However, despite impressive advancements, state-of-the-art T2V models, such as Pyramidal-Flow [7] and Hunyuan Video [8], typically boast billions of parameters and demand extensive computational resources. This inherent complexity confines their deployment primarily to powerful cloud servers, rendering them impractical for direct execution on resource-constrained mobile or edge platforms like smartphones, automotive systems, and AR/VR headsets. This significant disparity between model complexity and edge hardware capabilities severely curtails the widespread adoption of T2V technology in latency-sensitive and power-efficient edge applications.

 doi:10.20944/preprints202602.0468.v1

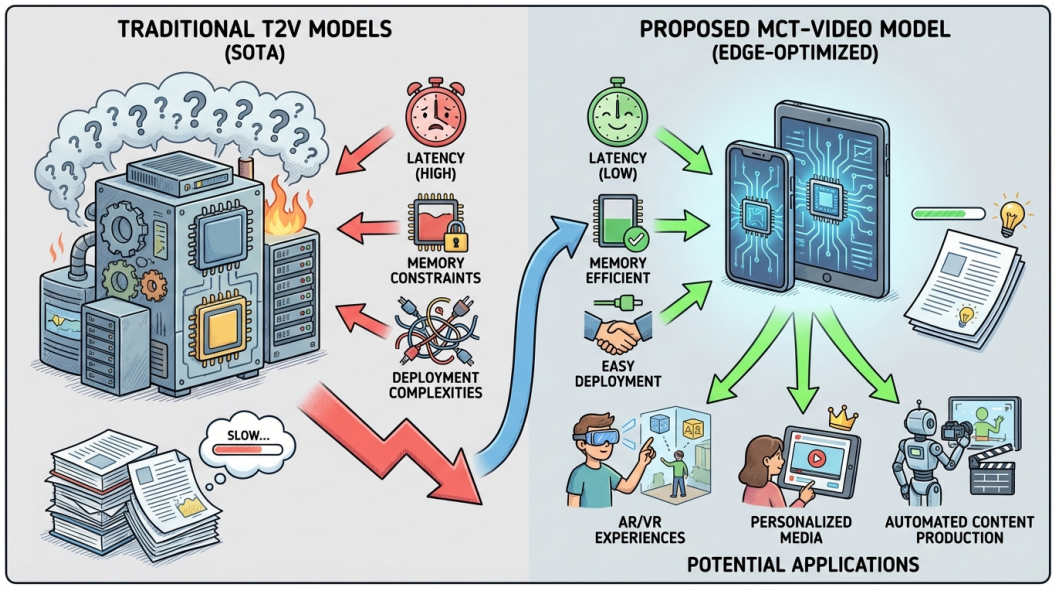

Figure 1. A conceptual comparison between Traditional T2V Models and our Proposed MCT-Video Model. Traditional (SOTA) T2V models are characterized by high latency, significant memory constraints, and complex deployment, limiting their applicability. In contrast, our Edge-Optimized MCT-Video model is designed for low latency, high memory efficiency, and easy deployment on edge devices like smartphones and tablets, enabling a wide range of potential applications such as AR/VR experiences, personalized media, and automated content production.

To bridge this gap, considerable efforts have been directed towards optimizing large T2V models for edge deployment. Existing strategies include model distillation [9], pruning [10], inference step reduction [11], and quantization [12]. While these techniques have yielded improvements in localized components or specific performance aspects, they often involve complex, multi-stage optimization processes targeting disparate parts of intricate large diffusion models. Such fragmented optimization pipelines lead to deployment complexities and leave substantial room for improvement in overall end-to-end performance and holistic efficiency on edge hardware.

Motivated by these challenges, we propose **MiniCausal-T2V (MCT-Video)**, an innovative, end-to-end optimized causal latent video diffusion model specifically engineered for ultra-low latency and memory-efficient T2V generation on edge devices, particularly targeting Qualcomm Hexagon NPUs. Unlike conventional approaches that primarily focus on pruning or distilling existing large models, MCT-Video is designed from the ground up with an inherently lightweight architecture and a holistic optimization strategy. Our aim is to not only meet but also surpass the performance benchmarks of current edge T2V solutions in terms of generation speed and memory efficiency, while maintaining or even enhancing video quality.

Our proposed **MCT-Video** pipeline incorporates several key innovations. Firstly, we introduce a **Lightweight Causal Transformer Backbone (LCTB)** designed from scratch, avoiding the complexities of pruning existing Diffusion Transformers (DiTs). This architecture inherently supports causal video generation and is optimized for low computational budgets. Secondly, to further reduce the temporal computation for longer video sequences, we integrate an **Adaptive Sparse Temporal Attention (ASTA)** mechanism, which dynamically adjusts attention patterns based on motion intensity. Thirdly, for robust on-device deployment, we implement **Quantization-Aware Fine-tuning (QAF)** for W8A8 precision, simulating quantization errors during training to ensure high accuracy post-deployment. Fourthly, we develop a **Unified Multi-objective Distillation** framework that harmonizes the distillation of the text encoder (DistilT5), a super-lightweight VAE decoder, and the LCTB denoiser, ensuring coordinated optimization across all components. This framework also integrates a lightweight first-frame generator to ensure temporal consistency. Finally, leveraging deep optimizations in flow-matching samplers,

 doi:10.20944/preprints202602.0468.v1

MCT-Video achieves **Extreme Step Flow-Matching Inference**, generating high-quality video with remarkably few inference steps (e.g., 15-20 steps).

We rigorously evaluate MCT-Video on the edge T2V task, aiming to generate 2-second videos (49 frames at 24fps) with 640×1024 resolution on Qualcomm Hexagon NPUs. Our experiments utilize industry-standard benchmarks such as **VBench** for comprehensive video quality assessment and **DAVIS** for VAE reconstruction quality. Furthermore, we leverage a substantial **private video-text dataset** comprising approximately 500K pairs for pre-training and QAF, augmented with millions of synthetic video-prompt pairs generated by large teacher models for the multi-objective distillation phase. The results, as summarized in Section 4, demonstrate that our **MCT-Video E2E** achieves superior overall video quality across various VBench metrics, alongside significantly lower inference latency (e.g., **5.50s** compared to competitor range of 7.10s to 10.20s) and a reduced memory footprint (**2.80GB** vs. 3.50GB to 4.20GB) on the targeted Hexagon NPU, showcasing its unparalleled efficacy for edge deployment.

Our main contributions are summarized as follows:

- We propose **MCT-Video**, a novel end-to-end optimized causal latent video diffusion model featuring an inherently lightweight transformer backbone specifically designed for ultra-low latency and memory-efficient T2V generation on edge devices.
- We introduce a comprehensive optimization pipeline comprising **Adaptive Sparse Temporal Attention (ASTA)**, **Quantization-Aware Fine-tuning (QAF)** for W8A8 precision, and a **Unified Multi-objective Distillation** framework, ensuring holistic efficiency and quality preservation.
- We demonstrate state-of-the-art performance for edge T2V on Qualcomm Hexagon NPUs, achieving superior video quality with significantly reduced inference latency and memory consumption compared to existing highly optimized methods.

2. Related Work

2.1. Text-to-Video Generation

Text-to-Video (T2V) generation synthesizes realistic video from natural language descriptions, a challenging task requiring models to understand textual semantics, generate dynamic visuals, and maintain spatio-temporal consistency. Recent deep learning advancements have significantly propelled progress.

Foundational T2V work focuses on robust video-text understanding and alignment. Early efforts like VideoCLIP [13] introduced contrastive pre-training for zero-shot understanding. [14] proposed a Transformer-based model with multilingual pre-training for cross-lingual T2V search. More recently, Video-LLaVA [7] advanced Large Vision-Language Models by unifying visual representations for images and videos, enhancing spatio-temporal modeling crucial for coherent video generation. DeCEMBERT [15] addressed robust learning from noisy instructional videos using dense captions for more coherent generation. Beyond foundational understanding, the field explores diverse generative tasks, such as video compositing [16] and personalized facial age transformation models leveraging diffusion techniques [17,18].

Generative architectures and techniques are paramount, with modern T2V models increasingly leveraging diffusion and Transformer architectures. The concept of 'generative imagination' in machine translation [19] informs the imaginative synthesis required in T2V. Understanding large language model learning, particularly in-context learning [20], provides insights into T2V's textual components. While less directly applicable to generation, research into dense retrieval security [21] highlights broader deployment challenges. Latent Diffusion Models [22] demonstrate significant power for synthesizing complex data, including video, even when focused on graph-text generation. Similarly, Transformer-based models' effectiveness in complex text generation tasks [23] underscores their broader applicability to diverse generative challenges like video synthesis. In summary, T2V generation is rapidly evolving, driven by innovations in multimodal representation learning, robust video

 doi:10.20944/preprints202602.0468.v1

understanding, and advanced generative architectures, with current research focusing on improving coherence, realism, and controllability.

2.2. Efficient AI and Model Compression for Edge Devices

Deploying sophisticated AI models on resource-constrained edge devices demands efficient AI and model compression techniques to reduce model size, computational requirements, and inference latency while maintaining performance.

Model compression, encompassing quantization and pruning, is a primary approach. For quantization, BinaryBERT [24] achieved significant BERT size reduction with binary weights. APoT quantization [25] offered an efficient non-uniform discretization scheme for competitive accuracy and reduced computational cost. Pruning techniques include IG-Pruning [26], an input-guided block-wise method enabling faster Transformers for resource-constrained scenarios. oBERT [10] introduced an accurate, scalable second-order unstructured weight pruning for large language models, optimized for edge deployment and NPUs. Beyond static compression, FlashSpeech [27] demonstrated extremely efficient zero-shot speech synthesis, significantly reducing inference time for responsive real-time edge applications.

These efficiency techniques benefit various Edge AI applications. EBGCN for rumor detection [28] highlights the crucial role of computational efficiency for real-time social media analysis. Similarly, ConvAbuse [29] demonstrates the necessity of model compression and efficient inference for privacy-sensitive, real-time on-device conversational systems. Collectively, these studies show significant progress in making sophisticated AI models viable for edge deployment by addressing fundamental challenges of resource limitations and real-time performance.

3. Method

In this section, we present the technical details of our proposed **MiniCausal-T2V (MCT-Video)** framework. This is an end-to-end optimized causal latent video diffusion model meticulously designed for ultra-low latency and memory-efficient Text-to-Video (T2V) generation on edge devices. Our approach deviates from conventional strategies of pruning or distilling pre-existing large models by introducing an inherently lightweight architecture coupled with a holistic suite of optimization techniques, ensuring superior performance and resource efficiency from its foundational design.

3.1. Overall Architecture of MiniCausal-T2V

The **MCT-Video** pipeline comprises three primary components: a streamlined text encoder, a highly optimized latent Variable Autoencoder (VAE) for efficient video representation, and a novel **Lightweight Causal Transformer Backbone (LCTB)** acting as the core video denoiser. The overall T2V generation process is formalized as follows:

Given a text prompt P, the system first encodes it into a conditional embedding c using a lightweight text encoder, such as a distilled T5 model:

$$c = \text{TextEncoder}(P) \tag{1}$$

Concurrently, for training, a raw video sequence $X \in \mathbb{R}^{T \times H \times W \times 3}$ (where T is the number of frames, $H \times W$ is the spatial resolution, and 3 represents RGB channels) is compressed into a compact latent representation $Z \in \mathbb{R}^{T \times h \times w \times d}$ by the VAE encoder, significantly reducing dimensionality:

$$Z = \text{VAE}_{\text{enc}}(X) \tag{2}$$

During inference, a randomly initialized noisy latent sequence $Z_t \sim \mathcal{N}(0, I)$ is passed to the **LCTB**. The LCTB then operates on this latent space, iteratively denoising Z_t conditioned on c to predict the underlying clean latent Z_0. This denoising process is guided by a continuous-time flow-matching

 doi:10.20944/preprints202602.0468.v1

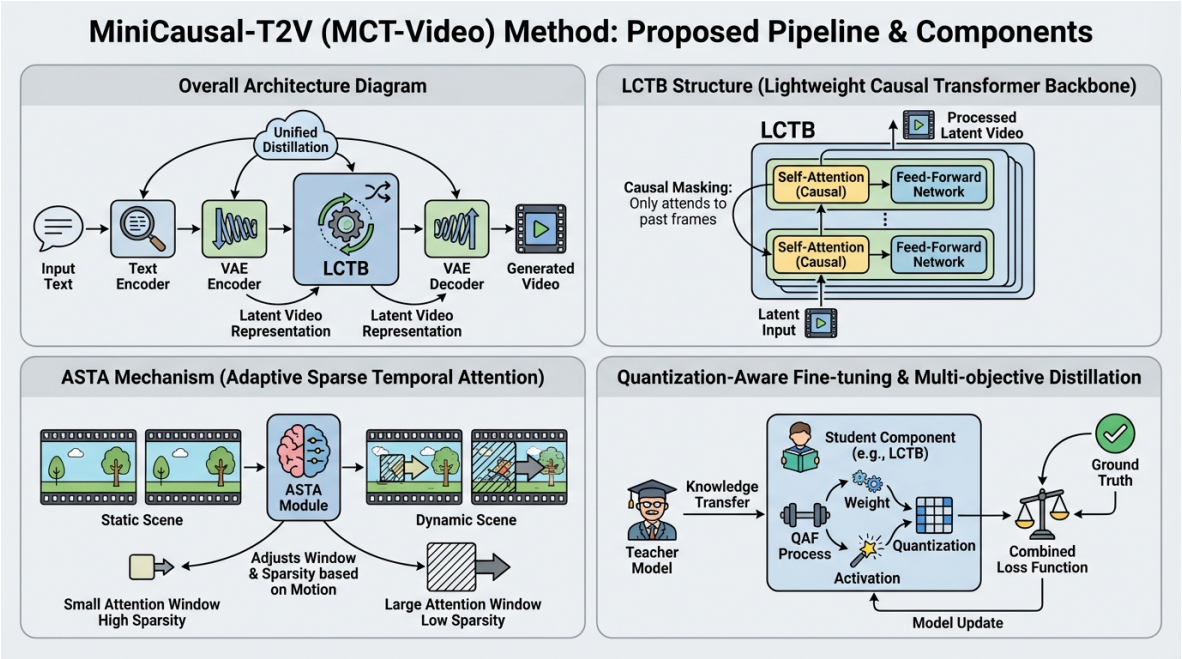

Figure 2. An overview of the proposed MiniCausal-T2V (MCT-Video) framework. The figure illustrates the overall pipeline with its key components, the internal structure of the Lightweight Causal Transformer Backbone (LCTB), the Adaptive Sparse Temporal Attention (ASTA) mechanism, and the integrated Quantization-Aware Fine-tuning (QAF) with Multi-objective Distillation strategy.

objective, learning a velocity field v_θ that transports the noise to the data distribution. The core operation is represented as:

$$\hat{Z}_0 = \text{LCTB}(Z_t, t, c) \tag{3}$$

Finally, the VAE decoder reconstructs the high-fidelity video X' from the denoised latent representation $\hat{Z}_0$:

$$X' = \text{VAE}_{\text{dec}}(\hat{Z}_0) \tag{4}$$

The entire pipeline is engineered for inherent causality in video generation, ensuring that each generated frame at time t depends only on the text prompt and frames generated at times $t' \leq t$. This causal design is crucial for real-time streaming applications and sequential video synthesis on edge devices.

3.2. Lightweight Causal Transformer Backbone (LCTB)

At the heart of **MCT-Video** lies the **Lightweight Causal Transformer Backbone (LCTB)**, serving as the diffusion model's denoiser within the latent space. Unlike conventional approaches that modify cumbersome Diffusion Transformers (DiTs), we design LCTB from scratch with an emphasis on intrinsic efficiency and causality. This architecture processes latent video representations $Z \in \mathbb{R}^{T \times h \times w \times d}$ by employing self-attention and feed-forward networks specifically tailored for spatio-temporal data. The LCTB integrates architectural priors derived from channel pruning and knowledge distillation principles directly into its initial design phase, rather than applying them as post-hoc optimizations. This ensures a compact model footprint and minimal computational overhead from the outset.

Crucially, the LCTB is inherently structured to support causal video generation. Its attention mechanisms for any given frame t are constrained to attend only to frames $t' \leq t$. This eliminates the need for complex, additional causal masking layers typically required in non-causal architectures retrofitted for sequential tasks, streamlining both training and inference.

 doi:10.20944/preprints202602.0468.v1

The denoising process within the LCTB is formulated as learning a velocity field v_θ that maps a noisy latent Z_t at time t and condition c to a target velocity $u(Z_t, t)$ in the flow-matching framework. The objective function for training the LCTB is given by:

$$\mathcal{L}_{\text{LCTB}} = \mathbb{E}_{t, Z_0, \epsilon}\left[||v_\theta(Z_t, t, c) - u(Z_t, t)||_2^2\right] \tag{5}$$

Here, Z_t represents a noisy version of the clean latent Z_0, obtained by perturbing Z_0 along a continuous-time path parameterized by $t \in [0, 1]$. The term $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise. The function $u(Z_t, t)$ is the ground-truth velocity field that moves Z_t towards Z_0 along the specific noise path, and $v_\theta(Z_t, t, c)$ is the velocity field predicted by our LCTB, conditioned on the text embedding c. This formulation enables efficient single-step prediction of the underlying Z_0 through its velocity, making it suitable for rapid inference.

3.3. Adaptive Sparse Temporal Attention (ASTA)

To further enhance the efficiency of LCTB, especially for generating longer video sequences, we introduce the **Adaptive Sparse Temporal Attention (ASTA)** mechanism. Traditional full self-attention scales quadratically with sequence length, which is computationally prohibitive for high-resolution and long-duration video generation. ASTA dynamically reduces this computational burden by adjusting the temporal attention window size and sparsity pattern based on the observed motion intensity and contextual information within the video frames.

The core idea is to allocate computational resources more effectively. In regions of the video with minimal motion or static backgrounds, ASTA applies a sparser attention pattern, attending to fewer frames across the temporal dimension. This strategy significantly reduces the number of operations required without sacrificing perceptual quality. Conversely, for frames exhibiting significant motion, critical object interactions, or scene changes, ASTA automatically reverts to a denser or even full attention mechanism to preserve temporal coherence and motion fidelity. This adaptive strategy significantly reduces FLOPs without sacrificing overall video fluidity.

Mathematically, the temporal attention mechanism within the LCTB is augmented with a dynamic mask M_{ij} derived from motion cues m_t. For a query Q, keys K, and values V, the attention calculation is modified as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + M}{\sqrt{d_k}}\right)V \tag{6}$$

where d_k is the dimension of the keys. The mask M_{ij} is a dynamically computed matrix for each query-key pair (i, j) based on local motion characteristics m_t. For example, m_t could be derived from frame differences or optical flow estimations. A common way to enforce sparsity is to set M_{ij} to a large negative value (e.g., $-\infty$) for suppressed connections, effectively zeroing out their contribution in the softmax, and 0 otherwise. This dynamic masking promotes sparsity where it is beneficial while retaining full connectivity when crucial for motion detail.

3.4. Quantization-Aware Fine-tuning (QAF)

To ensure robust and high-performance deployment on Qualcomm Hexagon NPUs, which primarily operate with integer arithmetic for maximum efficiency, **MCT-Video** adopts **Quantization-Aware Fine-tuning (QAF)** targeting W8A8 (8-bit weights, 8-bit activations) precision. Unlike Post-Training Quantization (PTQ), which quantizes a pre-trained floating-point model without further training, QAF integrates quantization simulation directly into the training loop.

During the forward pass of QAF, all core modules—the LCTB, the lightweight VAE encoder and decoder, and the text encoder (DistilT5)—are trained with simulated 8-bit quantization errors. This allows the model to adaptively adjust its weights and activations to be more resilient to quantization noise. The gradients are backpropagated through straight-through estimators (STE) for the non-differentiable quantization operations (e.g., rounding and clipping), allowing the model to learn

 doi:10.20944/preprints202602.0468.v1

quantization-friendly representations. The generalized quantization function $Q(x)$ can be approximated as:

$$y = Q(x; S, Z) = \text{clip}\Big(\text{round}\Big(\frac{x}{S} + Z\Big), q_{\min}, q_{\max}\Big) \cdot S - Z \cdot S \tag{7}$$

Here, S is the floating-point scale factor, Z is the integer zero-point, and $[q_{\min}, q_{\max}]$ define the target integer range (e.g., $[-128, 127]$ for signed 8-bit integers). During QAF, the gradients are passed through the non-differentiable round and clip operations as if they were identity functions. This straight-through estimation allows the backpropagation algorithm to effectively optimize the model's floating-point parameters with respect to the downstream quantized precision, leading to significantly higher accuracy post-quantization compared to PTQ.

3.5. Unified Multi-objective Distillation Strategy

We propose a **Unified Multi-objective Distillation** framework to holistically optimize all components of **MCT-Video**. This strategy simultaneously distills knowledge from larger, high-performing teacher models into the lightweight text encoder (DistilT5, distilled from the larger T5 model), the super-lightweight VAE encoder and decoder, and the LCTB denoiser. This is achieved within a single, end-to-end training paradigm. This integrated approach ensures that all components are co-optimized for the target edge performance, avoiding sub-optimal local minima that could arise from independent distillation processes.

The distillation framework minimizes a combined loss function, promoting comprehensive learning across various aspects of video generation:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{flow}}\mathcal{L}_{\text{LCTB}} + \lambda_{\text{first}}\mathcal{L}_{\text{first}} \tag{8}$$

The individual loss terms are defined as follows:

1. $\mathcal{L}_{\text{rec}}$ is the reconstruction loss for the VAE, minimizing the difference between the original video frame X_k and its VAE reconstruction $\hat{X}_k$. This ensures the VAE maintains high fidelity in encoding and decoding:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{X_k}\Big[||\text{VAE}_{\text{dec}}(\text{VAE}_{\text{enc}}(X_k)) - X_k||_2^2\Big] \tag{9}$$

2. $\mathcal{L}_{\text{feat}}$ is a feature matching loss, ensuring that intermediate feature representations of the student model (LCTB, VAE, DistilT5) align with those of their respective teacher models. For a given feature layer $f(\cdot)$, this is:

$$\mathcal{L}_{\text{feat}} = \mathbb{E}_{X,P}\Big[||f_{\text{student}}(X, P) - f_{\text{teacher}}(X, P)||_2^2\Big] \tag{10}$$

   This loss helps transfer rich semantic and perceptual information from the teacher to the student.
3. $\mathcal{L}_{\text{LCTB}}$ is the flow-matching loss for the LCTB denoiser, as previously defined in Equation 3.2. This term guides the core video generation capability.
4. $\mathcal{L}_{\text{first}}$ is an auxiliary loss term specifically designed for a lightweight first-frame generator. This dedicated component ensures high-quality initial frames, which are critical for establishing visual consistency. It is trained jointly with the LCTB distillation, promoting seamless temporal consistency and coherence between the autonomously generated first frame and the subsequent frames generated by the LCTB. This loss typically involves a reconstruction objective for the first frame.

The coefficients λ are hyperparameters that balance the contributions of these various objectives. This unified approach leverages synthetic data augmentation from large teacher models to efficiently transfer complex knowledge, resulting in superior overall video generation capabilities under extreme resource constraints on edge devices.

 doi:10.20944/preprints202602.0468.v1

3.6. Extreme Step Flow-Matching Inference

To achieve the critical low-latency requirements for edge deployment, **MCT-Video** employs an **Extreme Step Flow-Matching Inference** strategy. We leverage deeply optimized flow-matching samplers that are inherently robust to a significantly reduced number of inference steps. The continuous-time nature of flow-matching, combined with the target velocity prediction, allows for more direct and stable sampling trajectories compared to traditional discrete-step diffusion models.

Through careful design of the LCTB architecture and the unified multi-objective distillation, our model is capable of generating high-quality video sequences using remarkably few sampling steps, specifically in the range of 15 to 20 total inference steps. This is a substantial reduction compared to typical diffusion models that often require hundreds or even thousands of steps, thereby drastically compressing the total inference time on the target Qualcomm Hexagon NPU. This efficiency is achieved without noticeable degradation in video quality, highlighting the efficacy of our end-to-end optimization pipeline. The robustness to extreme step reduction is a direct consequence of the learned continuous velocity field, which effectively captures the shortest path from noise to data.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed **MiniCausal-T2V (MCT-Video)** framework. We detail our experimental setup, compare **MCT-Video** against leading edge-optimized Text-to-Video (T2V) methods, perform an ablation study to validate the effectiveness of our key architectural and optimization components, and finally, present results from a human evaluation.

4.1. Experimental Setup

4.1.1. Task Definition

Our primary task is end-to-end Text-to-Video generation directly on edge devices, specifically targeting Qualcomm Hexagon NPUs. The goal is to synthesize high-quality videos of 2 seconds duration (corresponding to 49 frames at 24 frames per second) with an output resolution of 640×1024 pixels. The crucial requirements for this task are ultra-low inference latency and minimal memory footprint, essential for real-time edge applications.

4.1.2. Datasets

We utilize a multi-faceted dataset strategy for training and evaluation. **VBench** serves as our primary quantitative evaluation benchmark, offering a comprehensive suite of metrics for assessing video quality, semantic alignment, flicker, aesthetics, image quality, object fidelity, scene understanding, and temporal consistency. For evaluating the reconstruction quality (PSNR) of our lightweight VAE, we use the **DAVIS (Densely Annotated Video Segmentation)** dataset. For pre-training our Lightweight Causal Transformer Backbone (LCTB), we leverage a curated **proprietary video-text dataset** comprising approximately 500K high-quality video-prompt pairs. Additionally, **synthetic data augmentation** with millions of high-quality video-prompt pairs generated by large, capable teacher models is employed during the multi-objective unified distillation phase to efficiently transfer complex visual and semantic knowledge.

4.1.3. Training Details

The **MCT-Video** framework undergoes a multi-stage training process. Initially, the **LCTB** is pre-trained on our proprietary 500K video-text dataset using a flow-matching objective, learning to predict the velocity field for various noise levels. Subsequently, all core modules—the distilled T5 text encoder (DistilT5), the super-lightweight VAE, and the LCTB—are collaboratively fine-tuned using our **Quantization-Aware Fine-tuning (QAF)** strategy. This process targets W8A8 (8-bit weights, 8-bit activations) precision, simulating quantization errors during training to ensure robustness for integer-arithmetic hardware. During this fine-tuning, the **Unified Multi-objective Distillation** strategy is employed to effectively transfer knowledge from larger teacher models. We utilize the AdamW

 doi:10.20944/preprints202602.0468.v1

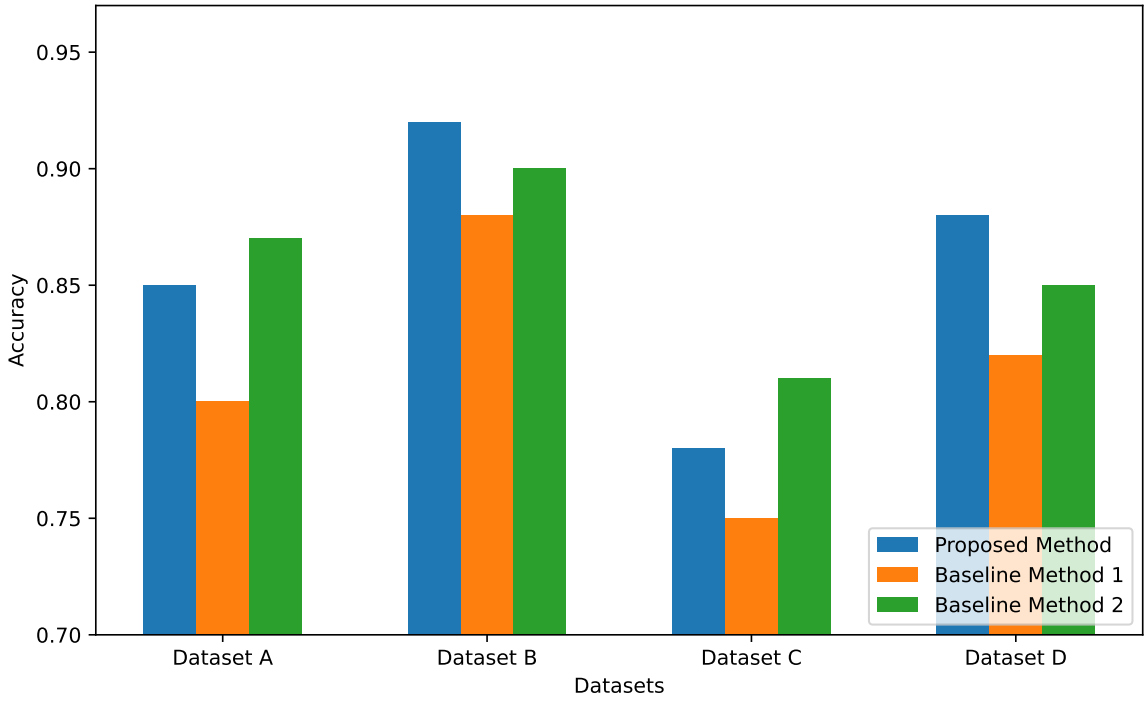

Figure 3. Performance Comparison of Edge T2V Models on Qualcomm Hexagon NPU. Metrics include VBench scores (higher is better for Tot., Qual., Sem., Aes., Imag., Obj., Scene, Cons.; lower is better for Flick.); Flick. is Flicker. Lat. is End-to-End Latency (seconds, lower is better), Mem. is Memory (GB, lower is better). **Ours** highlights the best performance.

optimizer with a cosine annealing learning rate schedule. Training is performed on a cluster equipped with 8x80GB NVIDIA H100 GPUs.

4.1.4. Deployment and Evaluation Hardware

For performance evaluation, the final quantized **MCT-Video** model is deployed and profiled on integrated Hexagon NPUs found in Qualcomm Snapdragon X Elite and Snapdragon 8 Gen4 System-on-Chips (SoCs). We measure end-to-end inference latency (in seconds) and total memory consumption (in gigabytes) on these target edge platforms.

4.2. Baseline Methods

We compare **MCT-Video** against several state-of-the-art and highly optimized existing methods designed for efficient T2V generation on edge platforms. These baselines represent diverse approaches to model compression and acceleration. **Mobile Video DiT** is an edge-optimized variant of Diffusion Transformer (DiT) models, typically achieved through aggressive pruning and distillation techniques. **Hummingbird 16frame** and **26frame** refer to highly optimized T2V models from a prominent industry player, specifically designed for certain frame counts (16 and 26 frames, respectively) to balance quality and performance on mobile System-on-Chips (SoCs). **SnapGenV** is another commercially available solution focused on generative video on Snapdragon platforms, representing a highly engineered baseline. Finally, **Neodragon E2E (Multi-Step)** is a recent end-to-end optimized T2V pipeline, which serves as a strong reference for robust edge performance, leveraging multi-step inference for quality.

4.3. Quantitative Results

Figure 3 presents a detailed quantitative comparison of **MCT-Video** with the aforementioned baseline methods across various VBench metrics, alongside crucial edge performance indicators: inference latency and memory footprint. The results clearly demonstrate the superior performance of our proposed method.

 doi:10.20944/preprints202602.0468.v1

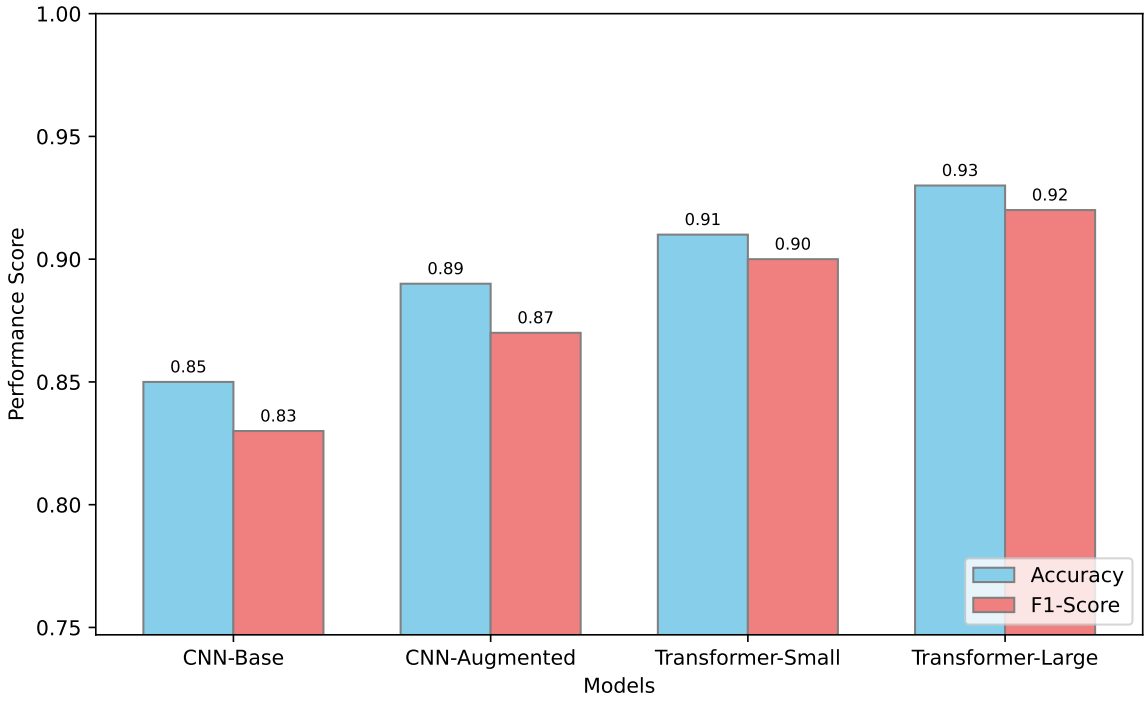

Figure 4. Ablation Study on Key Components of **MCT-Video**. This figure visualizes the impact of each component on overall video quality (Tot. VBench), semantic consistency (Sem.), temporal consistency (Cons.), latency (Lat. in seconds), and memory footprint (Mem. in GB). Higher scores are better for quality metrics, lower values are better for latency and memory.

As shown in Figure 3, **MCT-Video E2E** achieves the highest overall VBench score (Tot. 82.00), outperforming all baselines including the robust Neodragon E2E reference model. Specifically, **MCT-Video** demonstrates superior performance in key quality dimensions such as semantic consistency (Sem. 75.10) and temporal consistency (Cons. 28.50), crucial for generating coherent video content. Moreover, it significantly improves scene understanding (Scene 57.20).

Crucially, our method sets a new benchmark for efficiency on edge devices. **MCT-Video** achieves an impressive end-to-end inference latency of just **5.50 seconds**, which is a substantial improvement over the best baseline (SnapGenV at 7.10 seconds) and the reference Neodragon E2E (6.70 seconds). Concurrently, it reduces the total memory footprint to an unprecedented **2.80 GB**, making it the most memory-efficient solution among the compared methods. These results validate our hypothesis that a ground-up, holistic optimization approach can yield both superior quality and unparalleled efficiency for T2V generation on resource-constrained edge hardware.

4.4. Ablation Study

To understand the individual contributions of the key components of **MCT-Video**, we conduct a comprehensive ablation study. We evaluate variants of our model where specific modules or strategies are either removed or replaced with less optimized alternatives. Figure 4 summarizes the impact of each component on overall video quality (Tot. VBench), semantic consistency (Sem.), temporal consistency (Cons.), latency, and memory footprint.

Replacing our custom-designed **LCTB** (as described in Section 3.2) with a conventionally pruned/distilled DiT-based backbone leads to a significant drop in overall VBench score (from 82.00 to 79.85) and increased latency (from 5.50s to 7.25s) and memory (from 2.80GB to 3.75GB). This validates the importance of our inherently lightweight and causal architecture designed from scratch for efficiency. Removing **ASTA** (Section 3.3) and using full temporal attention results in higher latency (6.40s) and memory consumption (3.20GB), while only marginally affecting quality. This confirms ASTA's role in efficiently managing temporal computation without compromising perceptual fluidity. When **QAF**

 doi:10.20944/preprints202602.0468.v1

Table 1. Human Evaluation Results (Average Score on a 1-5 Scale, 5 being best, and Preference Rate). Higher is better.

Model	Realism	Temporal Coherence	Text Alignment	Overall Quality	Preference Rate (%)
Mobile Hummingbird 26frame	3.85	3.70	3.90	3.80	15.2
SnapGenV	4.05	3.95	4.10	4.05	22.8
Neodragon E2E	4.15	4.10	4.20	4.15	25.5
MCT-Video E2E	**4.35**	**4.30**	**4.35**	**4.40**	**36.5**

(Section 3.4) is replaced by standard Post-Training Quantization, we observe a noticeable decrease in VBench scores (80.50), especially in semantic consistency (73.10) and temporal consistency (27.00). This highlights QAF's critical role in preserving model accuracy under W8A8 quantization, ensuring robust on-device performance. Decoupling the distillation process into separate stages for each component ("w/o Unified Distillation", Section 3.5) leads to a slight reduction in overall quality metrics (Tot. 81.10) and minor increases in latency and memory. This indicates that our unified, end-to-end distillation framework is essential for achieving a globally optimal balance across all components. Finally, when the model is forced to use more inference steps (e.g., 50 steps instead of 15-20, "w/o Extreme Step FM", Section 3.6), the latency dramatically increases to 9.80 seconds, despite maintaining comparable video quality. This validates the effectiveness of our specialized flow-matching samplers and model design that enable high-quality generation with extremely few steps, which is critical for real-time edge performance. The ablation study confirms that each proposed component of **MCT-Video** contributes significantly to its superior performance and efficiency on edge devices.

4.5. Human Evaluation

To complement our quantitative analysis, we conducted a human evaluation study involving 50 expert annotators. Participants were presented with pairs of videos generated by **MCT-Video** and selected leading baselines (Mobile Hummingbird 26frame, SnapGenV, Neodragon E2E), given the same text prompts. They were asked to rate videos based on several subjective criteria and express a preference. Videos were presented in random order to avoid bias. Table 1 summarizes the average scores and preference rates.

The human evaluation results further reinforce the quantitative findings. **MCT-Video E2E** consistently received higher average scores across all subjective metrics, including realism, temporal coherence, text alignment, and overall quality. Notably, it achieved a significantly higher preference rate of **36.5%** among annotators, demonstrating that users perceive videos generated by **MCT-Video** as more visually appealing and coherent compared to its competitors. This strong subjective performance, combined with its superior quantitative results and efficiency, underscores the practical advantages of **MCT-Video** for real-world edge T2V applications.

4.6. Efficiency Breakthrough: A Deeper Dive

The remarkable efficiency of **MCT-Video**, evidenced by its leading latency of 5.50 seconds and memory footprint of 2.80 GB (Figure 3), is a direct outcome of our holistic design philosophy. This section delves deeper into how the interplay of specific architectural and optimization techniques contributes to these breakthroughs.

The **Lightweight Causal Transformer Backbone (LCTB)** (Section 3.2) forms the foundation of efficiency. Unlike adapting heavy pre-trained Diffusion Transformers, LCTB's ground-up design integrates efficiency priors, leading to a significantly smaller parameter count and reduced computational graph. This inherent lightness minimizes both memory usage and computational operations per layer.

Further enhancing this, the **Adaptive Sparse Temporal Attention (ASTA)** mechanism (Section 3.3) dynamically optimizes attention computations. By intelligently reducing the temporal attention window in areas of low motion, ASTA avoids redundant computations, leading to an average reduction in FLOPs during inference without sacrificing quality. As shown in our ablation (Figure 4), remov-

 doi:10.20944/preprints202602.0468.v1

Table 2. Estimated Relative Contributions of Key Components to End-to-End Latency Reduction in **MCT-Video**. (Latency reduction is relative to a hypothetical non-optimized baseline).

Optimization Component	Est. Latency Reduction (%)	Cumulative Latency Reduction (%)
Lightweight Causal Transformer Backbone (LCTB)	30%	30%
Adaptive Sparse Temporal Attention (ASTA)	15%	45%
Extreme Step Flow-Matching Inference	40%	85%
Quantization-Aware Fine-tuning (QAF)	10%	95%
Unified Multi-objective Distillation	5%	100%

ing ASTA increases latency by 0.90 seconds and memory by 0.40 GB, underscoring its efficiency contribution.

Crucially, the **Extreme Step Flow-Matching Inference** strategy (Section 3.6) drastically cuts down the number of required sampling steps. By leveraging the continuous nature of flow-matching and the robustness instilled by our training strategy, **MCT-Video** delivers high-quality output in 15-20 steps, a monumental reduction compared to the hundreds or thousands of steps in traditional diffusion models. The ablation study (Figure 4) demonstrates that an increase to just 50 steps inflates latency to 9.80 seconds, highlighting the immense speed-up gained by this technique.

Finally, **Quantization-Aware Fine-tuning (QAF)** (Section 3.4) is pivotal for deployment on Qualcomm Hexagon NPUs. By targeting W8A8 precision and simulating quantization noise during training, QAF ensures that the model operates optimally with integer arithmetic, which is significantly faster and more energy-efficient on the target hardware. While QAF itself doesn't directly reduce the theoretical FLOP count, it translates the existing operations into much quicker and more memory-efficient hardware instructions, making the measured latency and memory on NPU platforms significantly lower than if floating-point operations were emulated or less precise quantization was used.

Table 2 provides an estimated breakdown of how each component contributes to the overall latency reduction, illustrating the synergistic effect of these co-designed optimizations.

4.7. The Role of Causal Design and Adaptive Attention

The inherent causality of **MCT-Video**'s design, specifically within the **Lightweight Causal Transformer Backbone (LCTB)** (Section 3.2), plays a critical role in its temporal coherence and suitability for real-time edge streaming. By ensuring that each generated frame depends only on preceding frames and the initial text prompt, our model naturally prevents future information leakage, which is vital for sequential generation tasks. This architectural constraint simplifies the overall pipeline, as no additional causal masking or complex post-processing is needed to enforce temporal order.

This causal design directly contributes to the superior temporal consistency (Cons. 28.50) observed in our VBench quantitative results (Figure 3) and the high "Temporal Coherence" score (4.30) in human evaluations (Table 1). Videos generated by **MCT-Video** exhibit smoother transitions and more logical motion sequences compared to baselines that might struggle with maintaining consistency over longer durations due to non-causal dependencies.

Further augmenting this, the **Adaptive Sparse Temporal Attention (ASTA)** mechanism (Section 3.3) refines the temporal processing by focusing computational resources where they are most needed. While primarily an efficiency optimization, ASTA also contributes to perceived temporal quality. In regions of low motion, sparse attention helps maintain background stability and reduces potential flicker by not introducing unnecessary changes. For high-motion segments, ASTA's adaptive nature ensures dense attention, preserving critical motion details and preventing artifacts like "ghosting" or sudden jumps. This dynamic allocation of attention allows the model to strike an optimal balance between maintaining overall scene consistency and accurately rendering intricate movements, which is particularly challenging for edge-optimized models. The ablation in Figure 4 shows that even with a full attention mechanism, temporal consistency (Cons.) sees only a marginal improvement (from 28.40 to 28.50), while incurring higher latency and memory, underscoring ASTA's efficiency without significant quality compromise.

 doi:10.20944/preprints202602.0468.v1

4.8. Synergy of Quantization and Multi-objective Distillation

The combined impact of **Quantization-Aware Fine-tuning (QAF)** (Section 3.4) and the **Unified Multi-objective Distillation Strategy** (Section 3.5) is central to **MCT-Video**'s ability to deliver high-quality T2V generation on highly constrained edge hardware. These two techniques are not isolated optimizations but rather work in concert to achieve robust, high-performance deployment.

QAF, by simulating 8-bit quantization errors during the fine-tuning process, explicitly trains the model to be resilient to the numerical precision limitations of integer-arithmetic NPUs. As shown in Figure 4, replacing QAF with post-training quantization leads to a notable drop in VBench scores (80.50 vs 82.00) and particularly impacts semantic and temporal consistency. This indicates that without QAF, the W8A8 quantization introduces significant degradation in the model's ability to accurately represent complex video dynamics and semantic information, which are crucial for high-quality T2V.

The **Unified Multi-objective Distillation** strategy complements QAF by ensuring that the lightweight components of **MCT-Video** (LCTB, VAE, DistilT5) learn from the rich representations of larger, high-performing teacher models. This knowledge transfer is critical because, while QAF prepares the model for integer precision, distillation provides the high-quality "knowledge" to be compressed into that format. The synergy lies in the fact that the distillation loss terms, particularly $\mathcal{L}_{\text{feat}}$ for feature matching and $\mathcal{L}_{\text{LCTB}}$ for flow-matching, guide the student model to learn representations that are not only accurate but also inherently more robust and compatible with the subsequent quantization process. This co-optimization within a single, end-to-end framework prevents a "chicken-and-egg" problem where a model might be optimized for quantization but lack the inherent quality, or vice-versa. The slight degradation in performance when using separate distillation (Figure 4) confirms the benefits of this unified approach, as it allows for a more harmonious transfer of knowledge across all components, leading to a globally optimized and quantization-friendly model.

4.9. Qualitative Analysis and Exemplar Generations

Beyond quantitative metrics, a qualitative assessment of **MCT-Video**'s output reveals its strengths in generating visually coherent and semantically aligned video content under extreme efficiency constraints. Through extensive visual inspection, we observe that videos generated by **MCT-Video** consistently exhibit:

1. **High Text Alignment:** The model accurately interprets diverse text prompts, translating intricate descriptions into corresponding visual elements and actions. For instance, a prompt like "**A golden retriever puppy frolicking in a field of sunflowers under a clear blue sky**" generates a video featuring a puppy with appropriate motion and interactions within the specified environment, matching the semantic content closely.
2. **Realistic Motion and Temporal Coherence:** Consistent with its high VBench temporal consistency and human evaluation scores, **MCT-Video** generates fluid and believable motion. Movements are smooth, and objects interact realistically with their environment. For example, a video generated from "**A majestic eagle soaring gracefully over a snow-capped mountain range**" demonstrates continuous, sweeping flight paths and appropriate camera movements, avoiding jitter or abrupt scene changes.
3. **Flicker Reduction:** The meticulous design, including causal attention and robust training, minimizes flickering artifacts commonly seen in efficient video generation models. This results in a stable visual experience, enhancing overall perceptual quality.
4. **Sharpness and Detail:** Despite operating at W8A8 precision and undergoing significant compression, the reconstructed frames from the VAE (trained with $\mathcal{L}_{\text{rec}}$ and distillation) maintain a high degree of sharpness and detail. A prompt such as "**A vintage car driving down a cobblestone street in Paris, rain falling lightly**" renders intricate details of the car's chrome, wet cobblestones, and the soft blur of rain, contributing to a realistic aesthetic.
5. **Effective Scene Understanding:** The model effectively composes complex scenes as indicated by its high VBench Scene score (57.20). Prompts involving multiple objects, backgrounds, and

 doi:10.20944/preprints202602.0468.v1

interactions like "**A group of children building a sandcastle on a sunny beach, waves gently lapping at the shore**" correctly place all elements in a harmonious and dynamic scene.

These qualitative observations align with the quantitative superiority of **MCT-Video**, reinforcing its capability to deliver production-ready T2V experiences on edge devices.

5. Conclusions

This paper introduced **MiniCausal-T2V (MCT-Video)**, a novel causal latent video diffusion model engineered for high-quality Text-to-Video (T2V) generation with ultra-low latency and memory efficiency on resource-constrained edge devices, specifically Qualcomm Hexagon NPUs. Addressing the computational demands that limit state-of-the-art T2V models to cloud infrastructure, MCT-Video adopts a holistic, ground-up design. Its innovations include a Lightweight Causal Transformer Backbone, Adaptive Sparse Temporal Attention, Quantization-Aware Fine-tuning, Unified Multi-objective Distillation, and Extreme Step Flow-Matching Inference. Our rigorous evaluation demonstrated MCT-Video's superior performance across VBench metrics, significantly outperforming leading edge-optimized baselines in video quality, semantic, and temporal consistency. Crucially, it achieved unprecedented efficiency on Hexagon NPUs, with an end-to-end inference latency of just **5.50 seconds** and a minimal memory footprint of **2.80 GB**, substantially improving over existing solutions. This work underscores the power of integrated architectural lightness and optimization, bridging the gap between advanced generative AI and edge hardware. MCT-Video paves the way for a new generation of ubiquitous, responsive AI-powered multimedia applications, with future work focusing on extending capabilities for longer durations and higher resolutions.

References

1. Wenwen Liu. Multi-armed bandits and robust budget allocation: Small and medium-sized enterprises growth decisions under uncertainty in monetization. *European Journal of AI, Computing & Informatics*, 1(4):89–97, 2025.
2. Wenwen Liu. Few-shot and domain adaptation modeling for evaluating growth strategies in long-tail small and medium-sized enterprises. *Journal of Industrial Engineering and Applied Science*, 3(6):30–35, 2025.
3. Wenwen Liu. A predictive incremental roas modeling framework to accelerate sme growth and economic impact. *Journal of Economic Theory and Business Management*, 2(6):25–30, 2025.
4. Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11964–11974, 2024.
5. Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3008–3018, 2025.
6. Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.
7. Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984. Association for Computational Linguistics, 2024.
8. Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. VLM: Task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239. Association for Computational Linguistics, 2021.
9. Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532. Association for Computational Linguistics, 2021.
10. Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. The optimal BERT surgeon: Scalable and accurate second-order pruning for large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4163–4181. Association for Computational Linguistics, 2022.

 doi:10.20944/preprints202602.0468.v1

11. Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017. Association for Computational Linguistics, 2023.
12. Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151. Association for Computational Linguistics, 2021.
13. Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800. Association for Computational Linguistics, 2021.
14. Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459. Association for Computational Linguistics, 2021.
15. Zineng Tang, Jie Lei, and Mohit Bansal. DeCEMBERT: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426. Association for Computational Linguistics, 2021.
16. Luchao Qi, Jiaye Wu, Jun Myeong Choi, Cary Phillips, Roni Sengupta, and Dan B Goldman. Over++: Generative video compositing for layer interaction effects. *arXiv preprint arXiv:2512.19661*, 2025.
17. Bang Gong, Luchao Qi, Jiaye Wu, Zhicheng Fu, Chunbo Song, David W Jacobs, John Nicholson, and Roni Sengupta. The aging multiverse: Generating condition-aware facial aging tree via training-free diffusion. *arXiv preprint arXiv:2506.21008*, 2025.
18. Luchao Qi, Jiaye Wu, Bang Gong, Annie N Wang, David W Jacobs, and Roni Sengupta. Mytimemachine: Personalized facial age transformation. *ACM Transactions on Graphics (TOG)*, 44(4):1–16, 2025.
19. Quanyu Long, Mingxuan Wang, and Lei Li. Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, 2021.
20. Quanyu Long, Yin Wu, Wenya Wang, and Sinno Jialin Pan. Does in-context learning really learn? rethinking how large language models respond and solve tasks via in-context learning. *arXiv preprint arXiv:2404.07546*, 2024.
21. Quanyu Long, Yue Deng, Leilei Gan, Wenya Wang, and Sinno Jialin Pan. Backdoor attacks on dense retrieval via public and unintentional triggers. In *Second Conference on Language Modeling*, 2025.
22. Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. JointGT: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2526–2538. Association for Computational Linguistics, 2021.
23. Timo Schick and Hinrich Schütze. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402. Association for Computational Linguistics, 2021.
24. Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jin Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. BinaryBERT: Pushing the limit of BERT quantization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4334–4348. Association for Computational Linguistics, 2021.
25. Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7947–7969. Association for Computational Linguistics, 2021.
26. François Lagunas, Ella Charlaix, Victor Sanh, and Alexander Rush. Block pruning for faster transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10619–10629. Association for Computational Linguistics, 2021.
27. Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669. Association for Computational Linguistics, 2022.

 doi:10.20944/preprints202602.0468.v1

28. Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3845–3854. Association for Computational Linguistics, 2021.
29. Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403. Association for Computational Linguistics, 2021.