MMControl: Unified Multi-Modal Control for Joint Audio-Video Generation

Liyang Li*, Wen Wang*, Canyu Zhao, Tianjian Feng,
Zhiyue Zhao, Hao Chen, and Chunhua Shen†

Zhejiang University, Hangzhou, China

Abstract. Recent advances in Diffusion Transformers (DiTs) have enabled high-quality joint audio-video generation, producing videos with synchronized audio within a single model. However, existing controllable generation frameworks are typically restricted to video-only control. This restricts comprehensive controllability and often leads to suboptimal cross-modal alignment. To bridge this gap, we present **MMControl**, which enables users to perform **M**ulti-**M**odal **Control** in joint audio-video generation. MMControl introduces a dual-stream conditional injection mechanism. It incorporates both visual and acoustic control signals—including reference images, reference audio, depth maps, and pose sequences—into a joint generation process. These conditions are injected through bypass branches into a joint audio-video Diffusion Transformer, enabling the model to simultaneously generate identity-consistent video and timbre-consistent audio under structural constraints. Furthermore, we introduce modality-specific guidance scaling, which allows users to independently and dynamically adjust the influence strength of each visual and acoustic condition at inference time. Extensive experiments demonstrate that MMControl achieves fine-grained, composable control over character identity, voice timbre, body pose, and depth-guided scene structure in joint audio-video generation.

Keywords: Audio-Video Generation · Multimodal Controllable Generation · Cross-Modal Alignment

1 Introduction

The pursuit of photorealistic digital content creation has driven generative models to evolve from static image synthesis toward dynamic, high-fidelity video generation. With the advent of Diffusion Transformers (DiTs) [31], visual synthesis has achieved remarkable progress in both scalability and temporal consistency [23, 41]. More recently, the field has advanced toward a more ambitious frontier: native joint audio-video generation [13, 14, 30, 38]. Unlike conventional

* Equal contribution.
† Corresponding author.

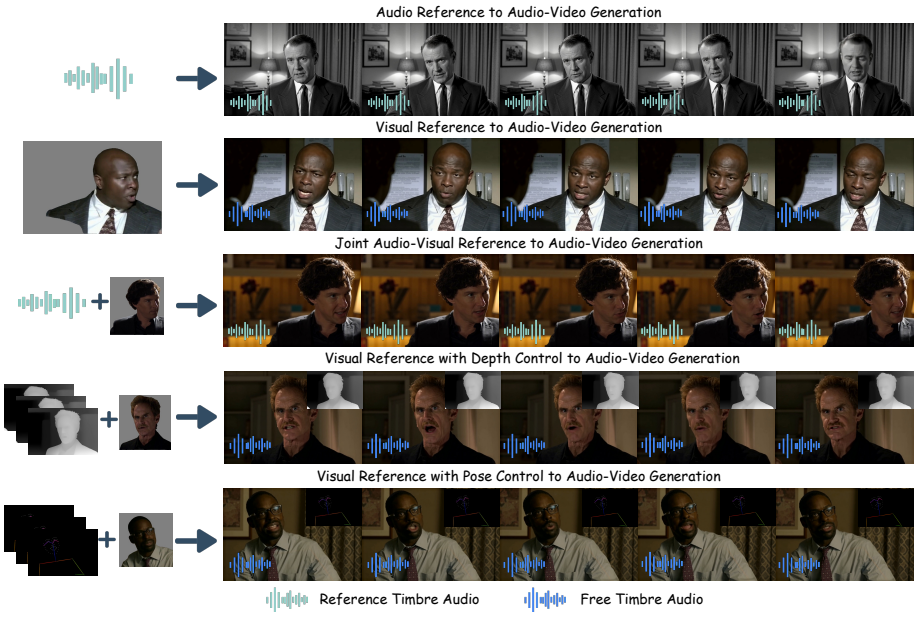

Fig. 1: Unified Multi-Modal Control results produced by MMControl. By providing different control signals, our model enables fine-grained control over joint audio-video synthesis. The teaser illustrates our model's ability to generate content that is: **Consistent** in voice and identity (via audio/visual references) and **Controllable** in structure and motion (via depth/pose guidance).

cascaded pipelines, these models learn to synthesize both modalities within a unified latent space, yielding intrinsic synchronization and a more holistic understanding of temporal dynamics.

Despite the impressive generation quality of recent joint DiT models [14, 38], they still lack fine-grained controllability over generated content. Existing methods are typically either visual-centric [49, 56] or audio-centric [18, 40, 46]. Visual-centric approaches primarily optimize visual control while under-modeling acoustic conditions, whereas audio-centric approaches emphasize speech or audio guidance but often struggle to produce naturally aligned audio-video sequences. These methods rely on decoupled controls across video and audio modalities, which makes coherent cross-modal alignment difficult. At the core lies the absence of a unified control mechanism that can simultaneously handle visual and acoustic conditioning signals. This exposes a critical gap: achieving precise multi-modal control within joint generation frameworks while preserving inherent cross-modal synchronization and collaborative priors remains highly challenging.

In this work, we focus on the task of Multi-Modal Controllable Joint Generation. The goal is to generate synchronized video and audio that follow textual descriptions while being conditioned on multimodal controls for visual struc-

ture (depth/pose), character identity, and vocal timbre. To this end, we propose `MMControl`, a new pipeline that introduces fine-grained controllability into joint audio-video generation. At the core of our method is a Multi-Modal Control Unit (MMCU), a unified interface that tokenizes heterogeneous inputs including reference images, reference audio clips, and depth/pose sequences, and aligns them into synchronized representations. This enables `MMControl` to handle these control scenarios under a single, consistent paradigm.

To integrate multimodal signals without compromising the pretrained generation capability of the backbone, we design a **Dual-Stream Bypass** architecture. This design introduces parallel, modality-specific trainable branches interleaved with frozen joint DiT layers. These branches extract contextual cues from the MMCU and inject them into the main diffusion process through gated residual connections. In addition, we introduce **Modality-Specific Guidance Scaling** at inference time, allowing users to independently adjust the influence of visual and acoustic controls. This flexibility enables `MMControl` to satisfy diverse creative requirements, ranging from strict structural adherence to expressive, identity-consistent synthesis. Extensive experiments demonstrate that `MMControl` establishes a new state-of-the-art in controllable joint generation, delivering exceptional visual fidelity, robust audio-video consistency, and precise structural alignment across reference-guided audio-video generation and depth/pose-controlled structural generation. Our main contributions are summarized as follows:

- We define and study the task of multi-modal controllable joint generation, providing a formal framework for jointly controlling identity, timbre, and depth/pose structure within a unified generative space.
- We propose `MMControl`, which introduces the **Multi-Modal Control Unit (MMCU)** to effectively unify and synchronize heterogeneous control signals as a versatile interface for joint diffusion.
- We design the **Dual-Stream Bypass architecture** for non-intrusive control-feature injection into frozen joint DiT layers, and further introduce Modality-Specific Guidance Scaling for flexible inference-time adjustment.

2 Related Work

Audio-Video Generation. Building on the success of image diffusion models [16, 35], video synthesis has evolved from foundational UNet-based architectures [1, 2, 17] to highly scalable DiT architectures [31]. Recent visual foundation frameworks [23, 41, 47, 54] pushed the boundaries of video synthesis. Beyond text-to-video synthesis, a significant body of work explores steering visual content via acoustic signals. This is most prominent in portrait animation, where pioneering methods like SadTalker [57] utilized 3D morphable models for motion mapping, while more recent diffusion-based methods [9, 10, 36, 37, 40, 46, 51] have achieved superior expressiveness and lip-sync accuracy. More recently, audio-to-video frameworks such as MoCha [45] and Wan2.2-S2V [41] have further

advanced the synthesis of high-fidelity video from acoustic signals. A transformative shift has recently occurred toward native joint audio-video generation, where both modalities are synthesized simultaneously within a single, unified framework. Pioneering industrial models such as Sora 2 [30] and Veo 3 [13] have demonstrated that modeling video and audio in a joint latent space leads to superior cross-modal synchronization and temporal realism. This paradigm is further advanced by frameworks like JavisDiT [25], LTX-2 [14], and Universe [42].

Controllable Video Generation. The quest for fine-grained controllability, which began with image generation, has remained a persistent mission that continues to drive innovation in the field of video generation. Pioneering works such as ControlNet [56] and T2I-Adapter [29] established the foundational paradigm for spatially-conditioned generation by integrating auxiliary bypass modules to process diverse geometric priors. These frameworks facilitate the precise steering of pretrained diffusion models through the incorporation of structural signals such as Canny edges, depth maps, and human poses. Building upon these advancements, GLIGEN [24] and Uni-ControlNet [59] further expanded the scope of controllability to encompass grounded generation and multi-modal conditional fusion. With the rapid advancement of video generation models, extending spatial control to the temporal domain has emerged as a critical research direction. Recent literature has predominantly focused on adapting image-based conditioning techniques to achieve motion-consistent video synthesis. A major line of work injects diverse structural signals and explicit motion priors (*e.g.*, motion vectors) to guide the global temporal generation process [7, 43, 58]. Building upon these foundations, subsequent approaches have further advanced toward fine-grained and interactive motion steering, enabling users to precisely dictate dynamic elements such as character poses or localized trajectories [11, 28]. Building upon these advancements, recent efforts [21,49] focus on developing unified frameworks for visual control. Despite these successes, existing methods primarily focus on the visual modality and do not account for the joint control of audio and video.

Customized Generation. Customized generation focuses on the preservation of user-specified subject identities. *Instance-specific customization* achieves high-fidelity identity preservation through dedicated per-subject fine-tuning. These optimization-based approaches effectively propagate subject features across temporal sequences and manage multiple identities via spatial masks and attention-based localization [4–6, 44, 48]. *End-to-end customization* provides a highly scalable, training-free alternative by integrating identity features through specialized encoder-based conditioning. By leveraging robust feature alignment and localized priors (*e.g.*, facial adapters), these methods seamlessly inject single or multiple identity representations in a single forward pass [12,15,19,26,55], thereby bypassing the need for tedious per-subject tuning. Despite rapid advancements in visual identity preservation, existing generative frameworks predominantly neglect the acoustic dimension. Specifically, the ability to control a character's unique voice timbre remains an under-explored frontier in joint audio-video generation. In this work, we explicitly address this critical gap.

3 Method

3.1 Task Definition

We define the task of **Multi-Modal Controllable Joint Generation**, aiming to synthesize temporally synchronized video $\nu \in \mathbb{R}^{t\times h\times w\times 3}$ and audio $\alpha \in \mathbb{R}^{t\times d}$ guided by a flexible combination of conditional signals. Generation is primarily driven by a textual prompt T following a structured format `[VISUAL]:` c_v `[SPEECH]:` c_s, where c_v provides the scene's semantic description and c_s specifies the spoken content to facilitate cross-modal semantic alignment. For fine-grained control, the model incorporates optional multi-modal signals: a reference image I^{ref} defining the character's visual identity, a reference audio clip A^{ref} providing the target voice timbre, and a sequence of structural constraints $S = \{s_1, s_2, \ldots, s_t\}$, instantiated as depth maps or pose sequences in this work. These composable conditions enable versatile scenarios ranging from basic text-to-AV synthesis to reference-guided and structure-guided joint generation. Formally, the process is formulated as:

$$\nu, \alpha = \mathcal{F}(T, I^{\text{ref}}, A^{\text{ref}}, S), \tag{1}$$

where $\mathcal{F}$ is a joint Diffusion Transformer mapping textual and multimodal inputs into temporally aligned streams, effectively ensuring identity consistency, timbre fidelity, and structural adherence to the provided depth or pose sequence.

3.2 Multi-Modal Control Unit (MMCU)

To bridge Section 3.1's raw control signals with the joint diffusion process, we introduce the **Multi-Modal Control Unit (MMCU)**. MMCU transforms diverse inputs into a synchronized representation $\mathcal{U} = \{T, V, A, M_v, M_a\}$, serving as our model's unified interface.

Textual guidance T derives from the described structured prompt. It is encoded into a joint embedding space via a pre-trained text encoder, providing global semantic context that guides visual and acoustic generation through cross-attention.

Visual control signal V and mask M_v are constructed by prepending the reference image latent to the structural guidance sequence. Letting $z_{img} = \text{VAE}(I^{\text{ref}})$ and $z_{s,i} = \text{VAE}(s_i)$ for $i \in \{1, \ldots, t\}$, the visual MMCU is formulated as:

$$\begin{aligned} V &= \{z_{img}\} + \{z_{s,1}, z_{s,2}, \ldots, z_{s,t}\}, \\ M_v &= \{0\} + \{1\} \times t, \end{aligned} \tag{2}$$

where $+$ denotes temporal concatenation. Binary mask M_v ensures the model preserves reference frame identity ($M_v = 0$) while following structural constraints S in generation frames ($M_v = 1$). For tasks without structural control, $\{z_{s,i}\}$ is replaced by empty latents $\{\mathbf{0}\} \times t$.

Similarly, acoustic control A and mask M_a align reference timbre with the target generation period. Let $\mathbf{z}_{aud} = \{z_{aud,1}, z_{aud,2}, \ldots,$

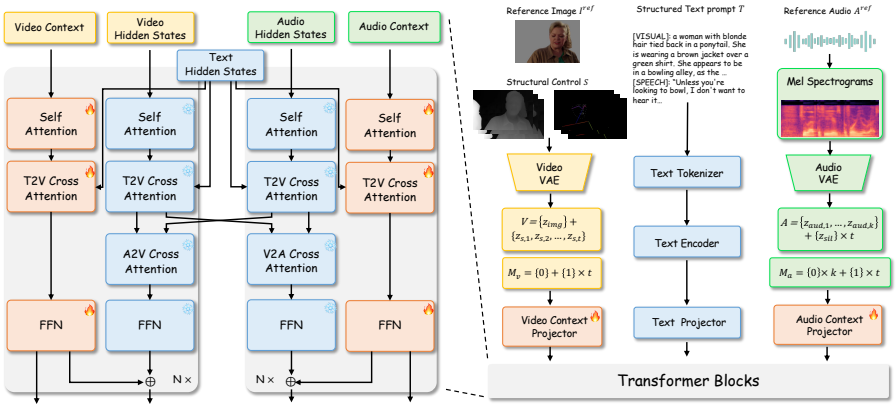

Fig. 2: Proposed Dual-Stream Bypass architecture for Joint DiT. The Dual-Stream Bypass mechanism introduces parallel, trainable visual and acoustic bypass branches interleaved with the frozen Joint DiT backbone at even-numbered layers.

$z_{aud,k}\}$ be the reference audio latent sequence from the Audio VAE, where k is the number of latent tokens for the reference duration. A sequence of silence latents $\mathbf{z}_{sil}$ of length t serves as a placeholder for generation. The acoustic MMCU is formulated as:

$$\begin{aligned} A &= \{z_{aud,1}, \ldots, z_{aud,k}\} + \{z_{sil,1}, \ldots, z_{sil,t}\}, \\ M_a &= \{\mathbf{0}\}_k + \{\mathbf{1}\}_t, \end{aligned} \tag{3}$$

where $\{\mathbf{0}\}_k$ and $\{\mathbf{1}\}_t$ are constant sequences of zeros and ones with lengths k and t. This ensures the model observes a reference audio segment to fix the speaker identity before the generative phase. While $\mathbf{z}_{sil}$ currently acts as this placeholder, this formulation allows for future extensions using specific acoustic controls like pitch or energy contours.

3.3 Architecture

Dual-Stream Bypass Architecture. To integrate MMCU signals while preserving pre-trained generative priors, we introduce the Dual-Stream Bypass mechanism. This architecture uses parallel trainable branches to encode multi-modal hints for targeted injection into the frozen Joint DiT backbone for precise conditional control.

Dual-Stream Context Projection. To effectively integrate MMCU signals, we propose a modality-specific projection mechanism. For each modality $m \in \{v, a\}$, the control sequence $\mathcal{U}_m$ and binary mask M_m are concatenated along the channel dimension:

$$\mathbf{z}_{in,m} = [\mathcal{U}_m; M_m], \tag{4}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. This formulation lets the model explicitly distinguish reference content from generative constraints via the aux-

iliary mask channel. A modality-specific Context Projector processes each fused sequence $\mathbf{z}_{in,m}$. To preserve generative priors, projector weights for latent features are inherited from the original backbone embedders, while mask channel weights are initialized from scratch to accommodate the new modality inputs.

Bypass Transformer Blocks. For modality-specific control, we introduce independent visual and acoustic bypass branches comprising Bypass Transformer Blocks, strategically interleaved with even-numbered backbone layers. To inherit base model capabilities, internal Self-Attention, Text Cross-Attention, and FFN layers are initialized using pre-trained weights. We deliberately omit explicit audio-to-video and video-to-audio cross-attention mechanisms within these branches to prioritize specialized intra-modality refinement and maintain high computational efficiency. At each even-numbered layer l, the bypass branch processes context tokens, with the resulting output mapped by a dedicated projection layer to form a contextual hint for residual injection into the main backbone. This output projector is zero-initialized to ensure training stability by preventing initial disturbances to the frozen pre-trained priors.

3.4 Inference Strategies for MMControl

To address multi-modal synchronization, we introduce modality-specific scaling alongside a two-stage refinement process, ensuring high-fidelity generation and flexible control over individual streams.

Modality-Specific Guidance Scaling. We propose Modality-Specific Guidance Scaling for decoupled control over distinct modalities. While bypass branches extract essential features, scaling factors γ_v and γ_a adaptively modulate their respective influence on joint generation for visual and acoustic streams. Specifically, in the l-th main transformer block, the hidden state $\mathbf{x}_{\text{main}}^{(l)}$ is updated via a gated residual connection that incorporates these contextual hints:

$$\mathbf{x}_{\text{main}}^{(l)} = \text{MainBlock}^{(l)}(\mathbf{x}_{\text{main}}^{(l-1)}) + \gamma_m \cdot \text{Hint}_m^{(l)}, \quad m \in \{v, a\}. \tag{5}$$

Here, $\text{Hint}_m^{(l)}$ denotes the modality-aware feature maps extracted from the corresponding l-th bypass block. Adjusting γ_v and γ_a during inference enables users to navigate the trade-off between strict conditional fidelity and generative creativity from pre-trained priors. For instance, a high γ_v value ensures rigorous adherence to complex pose sequences, whereas a moderate γ_a facilitates more natural prosody variations in synthesized speech.

Two-Stage Progressive Inference. Following the established LTX-2 [14] inference paradigm, we implement a Two-Stage Progressive Inference strategy to produce high-resolution outputs with optimal computational efficiency. This hierarchical framework explicitly decouples the initial formation of global semantic structures from subsequent refinement of fine-grained textures.

Stage 1: Semantic Base Generation. The model first generates a synchronized video and audio draft at a reduced spatial resolution ($h/2 \times w/2$). We use full MMCU conditions with a standard flow-matching schedule with Classifier-Free Guidance (CFG) to establish the foundational visual structure

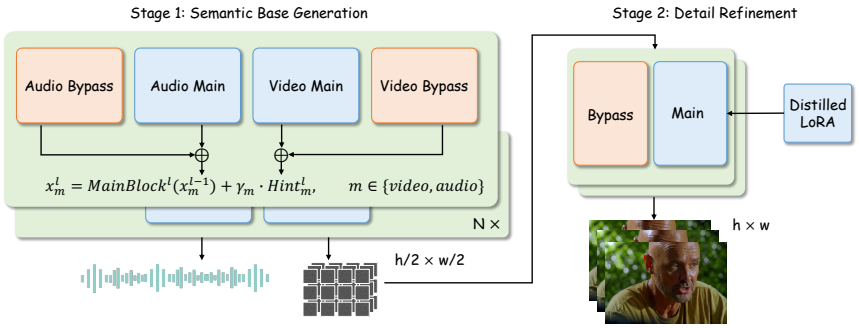

Fig. 3: Two-Stage Progressive Inference Strategy. Stage 1 forms a low-resolution $(h/2 \times w/2)$ semantic base with modality-specific scales (γ_v, γ_a) for independent visual and acoustic guidance. Stage 2 applies distilled LoRA for fine-grained refinement and upscaling to full resolution $(h \times w)$.

and acoustic timbre. During this phase, modality-specific scales γ_v and γ_a are employed to define the initial structural strength and cross-modal alignment.

Stage 2: Progressive Detail Refinement. This stage focuses on upscaling and refining the Stage 1 output to the target full resolution $(h \times w)$. A dedicated spatial latent upsampler first processes the low-resolution video latents. To enhance fine-grained detail synthesis without full denoising overhead, we integrate a frozen, pre-trained Distilled LoRA from LTX-2 into the backbone transformer.

During refinement, we apply a truncated noise schedule, typically spanning 3 steps, starting from the upsampled latents. CFG is disabled to preserve the established semantic layout, while MMCU signals are re-encoded at full resolution to provide precise guidance for high-frequency details. Finally, respective VAE decoders map these refined latents back into pixel space and high-fidelity audio waveforms.

4 Experiments

4.1 Experimental Setup

Data Preparation. `MMControl` is trained on a curated subset of 30,000 high-quality samples from Hallo3 [10]. The training objective covers three primary controllable tasks: generation conditioned on reference images and audio, reference images and depth maps, and reference images and poses, ensuring comprehensive multi-modal coverage. To maintain data integrity, we keep clips with a minimum duration of 2.0s, enabling non-overlapping reference-audio and target-video segments to prevent information leakage. During preprocessing, Grounded-SAM-2 [34] precisely segments characters for visual reference images. For structural guidance, Depth-Anything-v2 [52] computes depth maps, and DWPose [53] extracts whole-body poses. Qwen2.5-Omni [50] generates detailed captions for structured textual prompts to bridge the modality gap.

Implementation Details. `MMControl` is initialized from the pre-trained LTX-2 19B [14] backbone to leverage its generative knowledge. Optimization spans 7,200 steps using AdamW, with a peak learning rate of 1×10^{-5} regulated by a cosine annealing scheduler for smooth convergence. We train on four NVIDIA H200 GPUs with a per-GPU batch size of 2 and two gradient accumulation steps. To enable flexible composition and robust classifier-free guidance, we drop both visual and acoustic control signals with probability 0.1. Training takes approximately 12 hours, highlighting the efficiency of our bypass-based optimization.

Benchmark and Evaluation. To assess `MMControl`, we construct a comprehensive evaluation benchmark with 200 samples for comparison with audio-driven joint generation and another 200 for depth-conditioned generation. For general video quality assessment, covering motion smoothness, aesthetic appeal, and temporal consistency, we follow the multidimensional VBench [20] protocol. Task-specific evaluations use Sync-C and Sync-D [8, 32] to measure audio-visual synchronization precision. To quantify semantic alignment, Text CLIP similarity [33] is computed by averaging cosine similarity between the `[VISUAL]` portion of text prompts and eight uniformly sampled frames using a CLIP ViT-L/14 backbone. For identity preservation, Subject DINO similarity [3, 27, 34] is calculated between pre-segmented reference images and generated frames, leveraging a DINO ViT-B/16 model and Grounding-DINO with SAM2 for character isolation. We evaluate motion intensity via Dynamic Degree, which measures mean optical flow magnitude across 49 frames using RAFT-Large [39]. For structural control, we calculate Mean Absolute Error (MAE) between input depth maps and those extracted from synthesized videos to determine spatial fidelity.

4.2 Main Results

Table 1: Quantitative comparisons with prior methods. We report sync (Sync-C, Sync-D) and visual metrics; `MMControl` leads on most metrics, showing superior generation quality and audio-visual sync.

Method	Text CLIP Similarity↑	Subject DINO Similarity↑	Dynamic Degree↑	Aesthetic Quality↑	Motion Smoothness↑	Sync-D↓	Sync-C↑
SadTalker [57]	0.2410	0.8752	0.0057	0.5434	**0.9977**	10.696	2.474
AniPortrait [46]	0.2302	0.8857	0.0368	0.5350	0.9965	12.383	1.322
HunyuanCustom	0.2203	0.7816	0.6150	0.5003	0.9955	10.912	1.628
Hallo3 [10]	0.2337	0.8873	**0.9184**	0.5169	0.9667	**10.243**	2.550
`MMControl` **(Ours)**	**0.2546**	**0.8948**	0.5750	**0.5461**	0.9954	10.506	**2.716**

Comparison with Audio-Driven Baselines. Since customization frameworks such as Hallo3 [10], SadTalker [57], and AniPortrait [46] lack integrated audio-video synthesis, we benchmark them against `MMControl`. Following MoCha [45], we use the ground-truth first frame as a shared reference for all models, maintaining visual consistency. While baselines rely on ground-truth (GT) audio and

first frame, MMControl must jointly generate speech and video from text, a more complex generative task. Table 1 shows MMControl achieves the highest Sync-C score despite the baselines' GT audio advantage, proving our joint DiT captures inherent audio-visual correlations. MMControl leads in Text CLIP and Subject DINO similarity, confirming superior semantic alignment and identity preservation. Although Hallo3 exhibits a higher Dynamic Degree, MMControl better balances motion intensity and visual aesthetics, yielding superior appeal.

Structural Depth Control. We use depth maps as representative signals for structural control. As reported in Table 2, MMControl achieves a Mean MAE ($\times 100$) of **4.52**, substantially improving over VideoComposer (15.41) and recent SOTA VACE (5.35). Beyond structural precision, MMControl yields superior performance across semantic and motion metrics, especially subject similarity and dynamic degree. These results show that the Multi-Modal Control Unit (MMCU) preserves fine-grained spatial structures while maintaining high visual-semantic coherence during joint diffusion.

Table 2: Quantitative comparisons on structural depth control. We report visual metrics and depth MAE between input and generated videos. MMControl leads on most metrics and has the lowest structural error.

Method	Text CLIP Similarity↑	Subject DINO Similarity↑	Aesthetic Quality↑	Dynamic Degree↑	Motion Smoothness↑	Mean MAE ×100 ↓
VideoComposer [43]	0.2252	0.5888	0.4535	0.5900	0.9667	15.41
ControlVideo [58]	0.2450	0.6698	0.5652	0.2450	0.9852	10.64
VACE [21]	0.2434	0.8894	**0.5722**	0.4100	**0.9952**	5.35
MMControl **(Ours)**	**0.2453**	**0.8982**	0.5556	**0.6650**	**0.9952**	**4.52**

Structural Pose Control. As reported in Table 3, MMControl achieves the lowest pose MAE ($\times 100$) of **3.07**, outperforming VACE (3.29), Text2Video-Zero (6.40), and ControlVideo (6.59). It further obtains the best temporal consistency and motion smoothness, indicating the proposed control mechanism improves pose adherence while preserving stable temporal dynamics.

Table 3: Quantitative comparisons on structural pose control. We report visual metrics and pose MAE between input and generated videos. MMControl achieves the lowest pose error and best temporal consistency.

Method	Subject DINO Similarity↑	Dynamic Degree↑	Temporal Consistency↑	Motion Smoothness↑	Mean MAE ×100 ↓
Text2Video-Zero [22]	0.9823	**0.6750**	0.9682	0.9734	6.40
ControlVideo [58]	**0.9842**	0.2400	0.9823	0.9865	6.59
VACE [21]	0.9639	0.3800	0.9897	0.9946	3.29
MMControl **(Ours)**	0.9729	0.5250	**0.9906**	**0.9952**	**3.07**

Human and Audio Evaluation. We provide human evaluation and generated-audio quality analysis in the supplementary material. The user study shows `MMControl` obtains the highest overall score (3.58) across six perceptual criteria, while the audio evaluation reports speaker similarity (SIM-o 0.2574), naturalness (UTMOS 3.49), and transcription accuracy (7.96% weighted WER).

4.3 Qualitative Comparisons

Fig. 4: Qualitative comparison with baseline methods. As shown, baseline approaches such as HunyuanCustom, SadTalker, Hallo3 and AniPortrait often produce noticeable facial artifacts and largely static backgrounds, which limits scene expressiveness. `MMControl` generates high-fidelity videos with richer motion and natural transitions while maintaining stable character ID and precise audio-visual synchronization.

Comparison with Audio-Driven Baselines. As evidenced in Fig. 4, baseline methods such as AniPortrait and Hallo3 frequently exhibit noticeable facial artifacts and temporal jitters, especially during complex motion transitions. Furthermore, the backgrounds generated by these modular approaches remain largely static, an inherent limitation that suppresses scene dynamics and restricts overall narrative expressiveness. In contrast, `MMControl` produces high-fidelity videos with more nuanced motion and naturalistic temporal coherence that more faithfully align with the narrative prompts. Additional qualitative results are

provided in the supplementary material to further demonstrate the robustness of our framework.

Fig. 5: Qualitative comparison of depth control. The visual quality of images from VideoComposer and ControlVideo is lower. VACE shows weaker adherence to fine-grained depth cues (*e.g.*, hand gestures). Our method provides high visual quality and precise control adherence.

Comparison with Depth Control Baselines. The effectiveness of depth-control generation is further evaluated in Fig. 5. Compared to VideoComposer and ControlVideo, our approach exhibits significantly higher visual fidelity and subject consistency, avoiding the common identity-shifting artifacts and blurred textures seen in these baselines. While VACE produces high-quality frames, it shows weaker adherence to the intricate spatial structures and hand gestures provided in the depth control signal. In contrast, `MMControl` demonstrates stronger structural alignment, faithfully reconstructing complex motions while preserving the reference identity. This comparison focuses on visual structural adherence under the same depth input. Beyond this visual-control setting, `MMControl` further supports joint audio-video synthesis, enabling synchronized speech and lip motion within the same generation process.

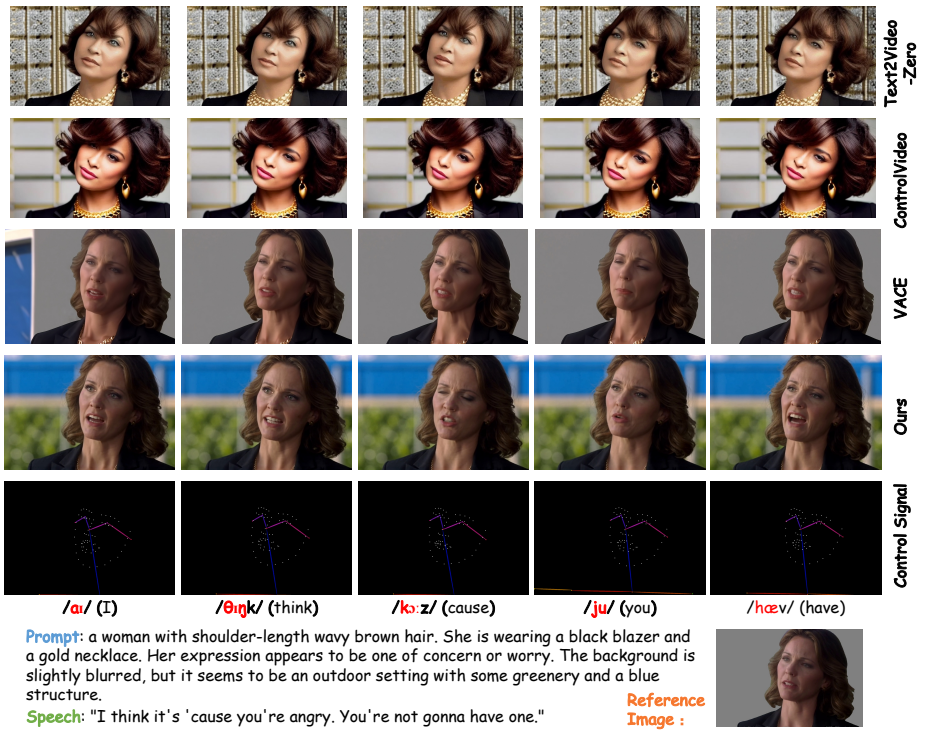

Fig. 6: Our method better adheres to structural pose signals and text prompts. Compared to baselines, `MMControl` generates videos with higher visual quality, more accurate identity, and precise audio-visual synchronization.

Comparison with Pose Control Baselines. The efficacy of pose-conditioned generation is qualitatively evaluated in Fig. 6. Compared to Text2Video-Zero and ControlVideo, the proposed approach shows superior visual fidelity and stronger adherence to the textual prompt, including the character's attire and outdoor background. While VACE achieves relatively high image quality, it is less consistent with the pose-conditioned scene composition in Fig. 6, including the specified background and character layout. In contrast, `MMControl` preserves the reference identity, follows the pose guidance, and generates synchronized audio in the same process, yielding a more coherent multi-modal output.

Decoupled Control Study. To assess independent control, we qualitatively vary the modality-specific scaling factors γ_v and γ_a during inference. In Fig. 7, when $\gamma_v = 0$ and $\gamma_a = 0$, the model generates a generic character and default voice guided exclusively by the textual prompt, reflecting the raw generative prior. With $\gamma_v = 1$, `MMControl` anchors the character's visual identity to the reference image while the acoustic timbre remains unconditioned. Conversely, with $\gamma_a = 1$ and $\gamma_v = 0$, the model preserves the reference voice timbre but synthesizes a novel visual persona. The full configuration $(\gamma_v = 1, \gamma_a = 1)$ achieves

Fig. 7: Modality-Specific Guidance Scaling. Varying γ_v and γ_a independently controls visual identity and acoustic timbre. With $\gamma_v = 1$, identity follows the reference image; with $\gamma_a = 1$, speech follows the reference voice.

precise synchronization of both identity and timbre. This decoupled control capability lets users flexibly navigate the trade-off between reference fidelity and generative creativity across distinct modalities.

5 Conclusion and Future Work

We study Multi-Modal Controllable Joint Generation for synchronized audio-video synthesis guided by textual, reference, and structural signals. MMControl is a unified framework built on a joint Diffusion Transformer. It uses the Multi-Modal Control Unit (MMCU) to harmonize inputs and a Dual-Stream Bypass to inject modality-specific hints into a frozen generative backbone. Modality-Specific Guidance Scaling enables independent control over visual identity and acoustic timbre during inference. Quantitative and qualitative results show that MMControl achieves state-of-the-art audio-visual synchronization and structural adherence to depth and pose constraints.

Despite promising results, challenges remain. Future work includes extending the framework to multi-character or conversational settings requiring complex cross-speaker synchronization and long-term temporal coherence. We also aim to explore end-to-end systems adapting to diverse control modalities and unseen speech styles without task-specific fine-tuning. This work lays a foundation for unified, user-centric multi-modal generation for personalized media creation.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chefer, H., Zada, S., Paiss, R., Ephrat, A., Tov, O., Rubinstein, M., Wolf, L., Dekel, T., Michaeli, T., Mosseri, I.: Still-moving: Customized video generation without customized video data. ACM Transactions on Graphics (TOG) **43**(6), 1–11 (2024)
5. Chen, H., Wang, X., Zhang, Y., Zhou, Y., Zhang, Z., Tang, S., Zhu, W.: Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3637–3646 (2024)
6. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Lee, K.S., Skorokhodov, I., Aberman, K., Zhu, J.Y., Yang, M.H., Tulyakov, S.: Multi-subject open-set personalization in video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6099–6110 (2025)
7. Chen, W., Ji, Y., Wu, J., Wu, H., Xie, P., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. arXiv preprint arXiv:2305.13840 (2023)
8. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian conference on computer vision. pp. 251–263. Springer (2016)
9. Cui, J., Li, H., Yao, Y., Zhu, H., Shang, H., Cheng, K., Zhou, H., Zhu, S., Wang, J.: Hallo2: Long-duration and high-resolution audio-driven portrait image animation. ICLR (2025)
10. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21086–21095 (2025)
11. Deng, Y., Wang, R., Zhang, Y., Tai, Y.W., Tang, C.K.: Dragvideo: Interactive drag-style video editing. In: European conference on computer vision. pp. 183–199. Springer (2024)
12. Fei, Z., Li, D., Qiu, D., Wang, J., Dou, Y., Wang, R., Xu, J., Fan, M., Chen, G., Li, Y., et al.: Skyreels-a2: Compose anything in video diffusion transformers. arXiv preprint arXiv:2504.02436 (2025)
13. GoogleDeepMind: Veo3. `https://deepmind.google/models/veo/` (2025), accessed: 28 June 2026
14. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
15. He, X., Liu, Q., Qian, S., Wang, X., Hu, T., Cao, K., Yan, K., Zhang, J.: Id-animator: Zero-shot identity-preserving human video generation. arXiv preprint arXiv:2404.15275 (2024)

16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
18. Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q., Lu, Q.: Hunyuancustom: A multimodal-driven architecture for customized video generation. arXiv preprint arXiv:2505.04512 (2025)
19. Huang, Y., Yuan, Z., Liu, Q., Wang, Q., Wang, X., Zhang, R., Wan, P., Zhang, D., Gai, K.: Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. arXiv preprint arXiv:2501.04698 (2025)
20. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
21. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
22. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15954–15964 (2023)
23. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
24. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
25. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Jiang, R., Luo, J., Fei, H., et al.: Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025)
26. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14951–14961 (2025)
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
28. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
29. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
30. OpenAI: Sora 2. `https://openai.com/index/sora-2/` (2025), accessed: 28 June 2026
31. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)

32. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia. pp. 484–492 (2020)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., Lu, J.: Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1982–1991 (2023)
37. Sun, X., Zhang, L., Zhu, H., Zhang, P., Zhang, B., Ji, X., Zhou, K., Gao, D., Bo, L., Cao, X.: Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. In: 2025 International Conference on 3D Vision (3DV). pp. 713–722. IEEE (2025)
38. Team, O., Yu, D., Chen, M., Chen, Q., Luo, Q., Wu, Q., Cheng, Q., Li, R., Liang, T., Zhang, W., et al.: Mova: Towards scalable and synchronized video-audio generation. arXiv preprint arXiv:2602.08794 (2026)
39. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
40. Tian, L., Wang, Q., Zhang, B., Bo, L.: Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: European Conference on Computer Vision. pp. 244–260. Springer (2024)
41. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
42. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025)
43. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36**, 7594–7611 (2023)
44. Wang, Z., Li, A., Zhu, L., Guo, Y., Dou, Q., Li, Z.: Customvideo: Customizing text-to-video generation with multiple subjects. IEEE Transactions on Multimedia (2026)
45. Wei, C., Sun, B., Ma, H., Hou, J., Juefei-Xu, F., He, Z., Dai, X., Zhang, L., Li, K., Hou, T., et al.: Mocha: Towards movie-grade talking character synthesis. arXiv preprint arXiv:2503.23307 (2025)
46. Wei, H., Yang, Z., Wang, Z.: Aniportrait: Audio-driven synthesis of photorealistic portrait animation. arXiv preprint arXiv:2403.17694 (2024)
47. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870 (2025)

48. Wu, T., Zhang, Y., Wang, X., Zhou, X., Zheng, G., Qi, Z., Shan, Y., Li, X.: Customcrafter: Customized video generation with preserving motion and concept composition abilities. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8469–8477 (2025)
49. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023)
50. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
51. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024)
52. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
53. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)
54. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
55. Yuan, S., Huang, J., He, X., Ge, Y., Shi, Y., Chen, L., Luo, J., Yuan, L.: Identity-preserving text-to-video generation by frequency decomposition. arXiv preprint arXiv:2411.17440 (2024)
56. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
57. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8652–8661 (2023)
58. Zhang, Y., Wei, Y., Jiang, D., ZHANG, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. In: The Twelfth International Conference on Learning Representations (2024)
59. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in neural information processing systems **36**, 11127–11150 (2023)