Baton: Explicit Semantic Blueprints for Joint Video-Audio Generation

Shuyuan Tu[1,*], **Qi Tian**[2,*], **Zihan Yang**[1], **Yue Wu**[2], **Xintong Han**[2],
Weijie Kong[2], **Jiangfeng Xiong**[2], **Jian-Wei Zhang**[2,†],
Zhao Zhong[2], **Liefeng Bo**[2], **Zuxuan Wu**[1,§], **Yu-Gang Jiang**[1]
[1]Institute of Trustworthy Embodied AI, Fudan University [2]Tencent Hunyuan
https://francis-rings.github.io/Baton

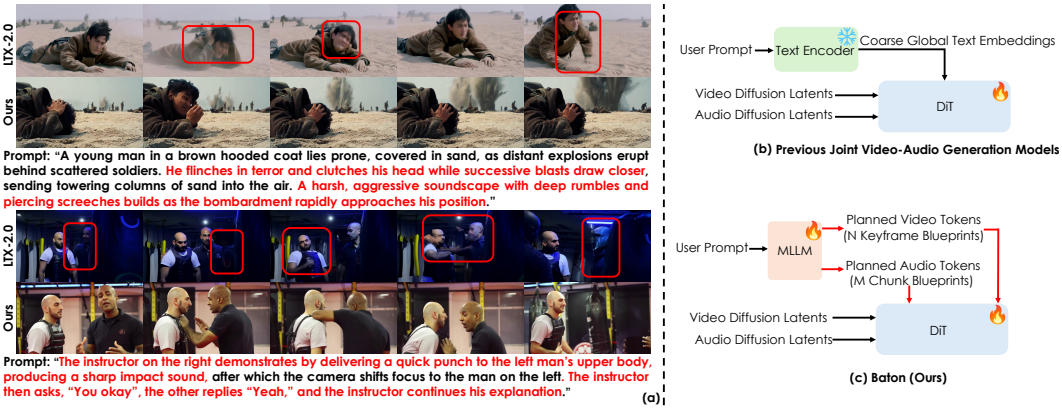


Figure 1: (a) Videos generated by Baton, showing its power to synthesize stable video-audio contents in semantically complex scenarios. LTX-2.0 is the latest joint video-audio model. (b) and (c) Framework comparison between previous models and Baton. Please refer to the demo for audio.

Abstract

Current open-source diffusion models struggle to generate stable and synchronized audio-visual content, particularly in scenarios demanding complex semantic reasoning. The root cause is that existing methods rely on coarse text embeddings from off-the-shelf encoders to guide audio-video denoising, which discards fine-grained semantics and, critically, lacks a shared long-horizon plan, leading to uncoordinated denoising trajectories and fragile cross-modal alignment. We propose Baton, the first framework that introduces explicit semantic planning into joint video-audio generation. Our key insight is that complementing coarse text guidance with semantically rich, modality-aware planned tokens, jointly reasoned and mutually aligned before denoising, can simultaneously restore fine-grained semantic detail and establish a shared blueprint that coordinates both audio and video denoising trajectories. Concretely, Baton first introduces the VA-Planner, a multimodal language model equipped with dual semantic alignment towers, where learnable queries cross-attend to both video and audio features to produce a pair of semantically aligned video and audio planned tokens as keyframe-level blueprints. These planned tokens are injected into the diffusion backbone via cross-attention layers, providing temporally grounded guidance complementary to coarse text embeddings. Since planned tokens do not share one-to-one spatial-temporal correspondence with diffusion latents, we further propose Relative Semantic RoPE, a relative positional encoding that maps planned tokens and latents into a shared spatial-temporal coordinate frame, enabling each latent to accurately attend to its

[*] Equal Contribution, [§] Corresponding Authors, [†] Project Leader.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

positionally corresponding semantic cues. Experiments on benchmarks show the effectiveness of Baton both qualitatively and quantitatively.

1 Introduction

Previous video generation models [4, 26, 38, 83] lack audio, limiting their real-world applicability. It drives joint video–audio generation to the mainstream in generative AI, as natively synchronized outputs are far more natural than video-first pipelines with post-hoc dubbing [28, 100]. However, current open-source models [100, 48, 21, 70, 63, 62] struggle to handle scenarios demanding complex semantic reasoning, such as compositional instructions with multi-stage actions or human–object interactions. It requires reasoning over long-range causal relationships, where generating stable video-audio remains significantly challenging.

While current open-source models [47, 46, 48, 21, 61, 85, 28, 100, 70] explore various strategies to enhance video-audio synchronization, they focus on simple prompts and use coarse text embeddings from off-the-shelf encoders to guide audio-video denoising, yet none of them design a module for cross-modal semantic comprehension that models how actions and causal intentions should coherently manifest and temporally progress across the visual and auditory streams (Fig. 1 (b/c)). Such global embeddings lack fine-grained understanding of how each event should jointly manifest across both modalities in a temporally coherent manner, causing the two denoising trajectories to drift apart and become adversarial, leading to severe cross-modal misalignment or distortion. Thus, these methods struggle to synthesize stable video-audio when confronted with semantically complex prompts. While some works [70, 21] apply LLMs [95, 3, 2] to enrich prompt details, they remain global embeddings injected into diffusion models [14, 24, 25, 67, 66, 60], which still cannot tackle the above issues.

In light of this, we propose Baton to perform semantic planning for stable joint video-audio generation, which explicitly disentangles semantic reasoning and synthesis (Fig. 2). In particular, Baton first introduces the VA-Planner, which is a learnable multimodal language model (MLLM) with dual semantic alignment towers to conduct semantic reasoning. Each tower uses learnable queries that cross-attend to both video and audio features, acting as modality bridges to fuse semantic cues and capture cross-modal correspondence. Unlike global text embeddings, VA-Planner reasons over the user prompts to produce a pair of semantically aligned video and audio planned tokens. These tokens, coordinated via the above towers, serve as keyframe-level blueprints that encode how visual scenes and their accompanying sounds should jointly unfold over time, providing fine-grained, modality-aware guidance that global text embeddings cannot offer. The planned tokens are then fed into the diffusion model via cross-attention, grounding each denoising step with temporally structured semantic cues. As cross-modal semantic coordination is explicitly established before denoising, both the video and audio generation trajectories are anchored to a shared, pre-aligned semantic roadmap. It prevents the two modalities from drifting into adversarial dynamics, as both trajectories are guided to evolve along semantically coherent paths rather than blindly following vague global signals.

However, since planned tokens and diffusion latents do not share one-to-one spatial-temporal correspondence, Baton introduces Relative Semantic RoPE (RS-RoPE), a relative positional encoding that maps them into a shared coordinate frame so that each latent can accurately attend to its positionally corresponding semantic guidance. Thus, Baton transforms the generation from blindly denoising under vague global signals to following a semantically grounded and cross-modally coordinated plan, yielding stable video-audio content even for prompts requiring semantic reasoning.

As shown in Fig. 1 (a), while LTX-2 [21] suffers from body distortion and misalignment between audio and video, Baton can accurately synthesize stable video-audio content based on user prompts, even in long-horizon scenarios involving multiple actions and complex human–object interactions.

In conclusion, our contributions are as follows: (1) We propose the novel VA-Planner to perform semantic reasoning over user prompts to produce a pair of semantically aligned video and audio planned tokens as keyframe-level blueprints, anchoring both denoising trajectories to a shared, coordinated semantic plan. To our knowledge, Baton is the first to disentangle semantic planning with synthesis for joint video-audio generation. (2) We introduce the RS-RoPE, which maps planned tokens and diffusion latents into a shared spatial-temporal coordination, enabling each latent to attend to its corresponding semantic cues. (3) Experiments show the superiority of Baton over the SOTA.

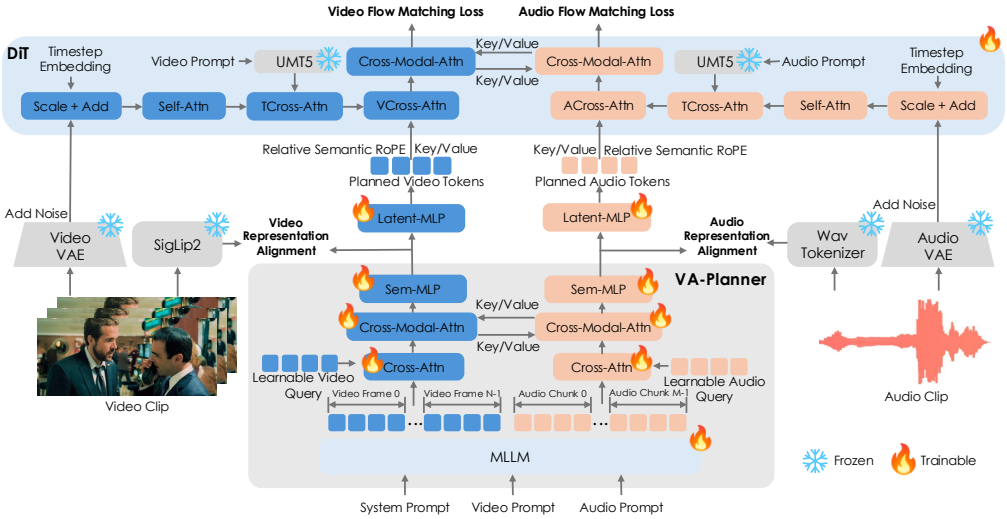


Figure 2: Architecture of Baton. Given the user prompts, Baton utilizes an MLLM to perform semantic reasoning and predict planned video/audio tokens, which are fed to the DiT via cross-attention for offering keyframe-level detailed blueprints during joint video-audio denoising.

2 Related Work

Video Generation. The capacity of high fidelity and diversity in diffusion models [14, 24, 25, 74, 76, 53, 67, 66, 60, 52, 23, 82, 97, 41] has promoted dramatic interest in their application for video generation [12, 35, 39, 16, 94, 96, 81, 77, 78, 29]. Early video diffusion models [65, 5, 20, 89, 86, 6, 73, 75, 80] mostly utilize the U-Net for video generation by inserting temporal layers into the backbone. Recent works [4, 26, 38, 83] replace the U-Net with the Diffusion-in-Transformer (DiT) [55] for scalability. However, the above models only generate silent videos, restricting their practical applications.

Joint Video-Audio Generation. The high fidelity of industry-leading models (Veo 3 [18] and Seedance 2.0 [7]) sparks the research interest in joint video-audio generation [47, 46, 48, 21, 61, 85, 28, 100, 70, 79, 31, 30]. JavisDiT [47] and UniAVGen [100] focus on joint video-speech generation. UniVerse-1 [85] and Ovi [48] can also synthesize ambient sounds. Harmony [28] enhances the synchronization via cross-task synergy training. MOVA [70] and LTX-2 [21] improve the quality by scaling the model and data. JavisGPT [46] utilizes unified tokens to guide the generation. However, the above models mostly apply frozen language models (T5 [59] or Qwen [95]) to extract the global text embeddings and inject them into DiT. Thus, they struggle to synthesize stable audio-visual content in semantically complex scenarios. By contrast, Baton addresses these issues.

Unified MultiModal Generation. Some works [8, 13, 22, 37, 43, 44, 50, 57, 69, 84, 90, 91, 92, 98, 101, 102] tend to achieve multimodal understanding and generation via a single framework. Early works [37, 84, 87, 98, 22] convert pretrained semantic encoders [72] to text-aligned visual encoders for generation. Recent image generation models [9, 17, 13, 91, 102, 8, 54] use MLLMs to predict visual tokens, coupled with diffusion heads for image synthesis. However, these methods primarily focus on static image generation, with limited exploration of joint video–audio generation under temporal semantic reasoning, such as multi-stage actions or complex human–object interactions.

3 Method

Shown in Fig. 2, Baton explicitly disentangles semantic reasoning and synthesis for establishing a modality-aware blueprint that coordinates both audio and video denoising. Concretely, the prompts are first fed to an MLLM to perform semantic reasoning and predict a pair of video/audio planned tokens as shared blueprints, as detailed in Sec. 3.1. To provide blueprint guidance, the planned tokens are further injected into a DiT via cross-attention. The DiT has the same dual-branch architecture as Ovi [48]. Notably, as the planned tokens and the diffusion latents live on misaligned spatio-temporal grids, we introduce Relative Semantic RoPE to resolve the mismatch, as detailed in Sec. 3.2.

3.1 VA-Planner

Current joint video-audio generation models [21, 70, 100, 47, 28] only rely on global text embeddings gained from a frozen LLM [59, 95] to guide the generation. It encodes the entire prompt into a vague embedding without decomposing it into modality-specific temporal semantics or modeling how visual events and auditory cues should correspond at each generation stage, leaving both denoising branches to independently interpret the vague signal and inevitably diverge under complex scenarios. To tackle this, we propose the VA-Planner, which uses a trainable MLLM to perform semantic reasoning and predict modality-specific yet mutually aligned planned tokens. Each token encodes a localized semantic context specifying what occurs, where, and when, while the paired video and audio tokens are jointly reasoned in a single autoregressive pass to ensure cross-modal consistency at every time point. Both generation trajectories are anchored to a shared semantic roadmap before denoising, preventing the two modalities from drifting into adversarial dynamics. It can restore fine-grained semantics and establish the shared cross-modal blueprints that global embeddings cannot provide.

Concretely, to plan a video with N keyframes (sampled at FPS $= 6$) and M audio chunks (each chunk covers one second of audio), we construct a structured user prompt T_{user} (Appx. A.6) that concatenates the system prompt T_{sys}, video prompt T_v, audio prompt T_a, and special delimiter tokens (T_v^{tag} and T_a^{tag}) that delineate the visual and auditory planning regions:

$$T_{\text{user}} = [T_{\text{sys}};\ T_v;\ T_a;\ T_v^{tag};\ T_a^{tag}], \tag{1}$$

where T_v^{tag} = <|v_start|>;<|img_pad|>$\cdots$<|img_pad|>;<|v_end|> and T_a^{tag} = <|a_start|>;<|aud_pad|>$\cdots$<|aud_pad|>;<|a_end|>. Each <|img_pad|> is a placeholder for one visual semantic token, with n_v tokens per keyframe, yielding $L_v = N \times n_v$ video tokens in total. Since predicting visual tokens for all frames incurs prohibitive cost, we sample N keyframes at FPS $= 6$ instead of all video frames. Each <|aud_pad|> is a placeholder for one audio semantic token, with n_a tokens per chunk, yielding $L_a = M \times n_a$ audio tokens. Since $n_a << n_v$, predicting all audio chunks is affordable. We adopt the same multimodal RoPE as Qwen3-VL [2] in the MLLM to jointly encode the heterogeneous positions of text, video, and audio tokens in T_{user}. To encode how the described visual scenes and accompanying sounds should semantically unfold over time, the MLLM performs autoregressive reasoning over T_{user}, and we extract the hidden states at the pad token positions to gain the video and audio hidden states:

$$\boldsymbol{H}_v = \texttt{MLLM}(T_{user})_{\texttt{[v_start+1:v_end]}} \in \mathbb{R}^{L_v \times D}, \quad \boldsymbol{H}_a = \texttt{MLLM}(T_{user})_{\texttt{[a_start+1:a_end]}} \in \mathbb{R}^{L_a \times D}, \tag{2}$$

where D is the hidden dimension. By placing the video and audio planning regions after all text prompts within a single autoregressive sequence, the causal attention of MLLM naturally allows both $\boldsymbol{H}_v$ and $\boldsymbol{H}_a$ to condition on the full prompt context, while $\boldsymbol{H}_a$ further attends to the preceding $\boldsymbol{H}_v$, establishing an implicit cross-modal dependency at the reasoning stage itself.

Since the planned tokens aim to encode concrete perceptual structures rather than remain in the MLLM's language-centric space, we introduce dual semantic alignment towers that project them into the continuous feature domains of pretrained perceptual encoders (SigLip2 [72] for video and WavTokenizer [34] for audio). As the MLLM's causal dependency is unidirectional ($\boldsymbol{H}_v$ cannot attend to $\boldsymbol{H}_a$), the video plan is blind to audio cues. The towers address this via bidirectional cross-modal attention. Each tower employs learnable queries, allowing them to flexibly distill the most relevant semantic cues from $\boldsymbol{H}_v$ and $\boldsymbol{H}_a$ into planned tokens. For the video tower, the learnable query $\boldsymbol{Q}_v \in \mathbb{R}^{L_v \times D}$ first cross-attends to $\boldsymbol{H}_v$ to extract video-specific semantics, then performs cross-modal attention $\texttt{CMAttn}_{a\to v}(\cdot)$ with $\boldsymbol{H}_a$ to absorb complementary auditory cues, and is projected to the target perceptual encoder dimension (SigLip2 or WavTokenizer) via a Sem-MLP $\texttt{SMLP}_{v/a}(\cdot)$:

$$\boldsymbol{H}_v^{sem} = \texttt{SMLP}_v(\texttt{CMAttn}_{a\to v}(\texttt{CAttn}_v(\boldsymbol{Q}_v, \boldsymbol{H}_v), \boldsymbol{H}_a)), \tag{3}$$

where $\texttt{CAttn}_v(x, y)$ and $\texttt{CMAttn}_{a\to v}(x, y)$ refer to cross-attention with x as Query and y as Key/Value. Symmetrically, the audio tower produces:

$$\boldsymbol{H}_a^{sem} = \texttt{SMLP}_a(\texttt{CMAttn}_{v\to a}(\texttt{CAttn}_a(\boldsymbol{Q}_a, \boldsymbol{H}_a), \boldsymbol{H}_v)), \tag{4}$$

where $\boldsymbol{Q}_a$ is the learnable query. Since $\boldsymbol{H}_v$ and $\boldsymbol{H}_a$ have different temporal densities, we apply timestamp-based RoPE in $\texttt{CMAttn}_{v\to a}(\cdot)$ and $\texttt{CMAttn}_{a\to v}(\cdot)$ to map both modalities onto a shared time axis, as detailed in Appx. A.3. Due to the towers, $\boldsymbol{H}_v^{sem}$ and $\boldsymbol{H}_a^{sem}$ both encode a mutually consistent temporal blueprint rather than two independent plans. Notably, Baton employs two distinct RoPE designs at different stages: (i) timestamp-based RoPE within the dual towers' `CMAttn` for aligning cross-modal tokens during planning (Appx. A.3), and (ii) Relative Semantic RoPE in the DiT's `VCAttn/ACAttn` for aligning planned tokens with diffusion latents during denoising (Sec. 3.2).

3.2 Relative Semantic RoPE

To provide keyframe-level blueprints during denoising, we inject $\boldsymbol{H}_v^{sem}$ and $\boldsymbol{H}_a^{sem}$ into dual-branch-based DiT via cross attention. However, the video/audio diffusion latents and $\boldsymbol{H}_v^{sem}$/$\boldsymbol{H}_a^{sem}$ do not share one-to-one spatiotemporal correspondence (Fig. 4) due to mismatched compression rates. Latents attend uniformly to all planned tokens by content similarity alone, collapsing the structured plan into a vague global signal. To address this, we propose the Relative Semantic RoPE (RS-RoPE).

In particular, we first employ Latent-MLP $\texttt{LMLP}_{v/a}(\cdot)$ to project planned tokens $\boldsymbol{H}_v^{sem}$ and $\boldsymbol{H}_a^{sem}$ from the spaces of SigLip2 and WavTokenizer to the diffusion latent dimension d:

$$\boldsymbol{H}_v^{dit} = \texttt{LMLP}_v(\boldsymbol{H}_v^{sem}) \in \mathbb{R}^{L_v \times d}, \quad \boldsymbol{H}_a^{dit} = \texttt{LMLP}_a(\boldsymbol{H}_a^{sem}) \in \mathbb{R}^{L_a \times d}, \tag{5}$$

$\boldsymbol{H}_v^{dit}$ and $\boldsymbol{H}_a^{dit}$ subsequently perform VCross-Attn $\texttt{VCAttn}(\cdot)$ and ACross-Attn $\texttt{ACAttn}(\cdot)$ with diffusion latents. However, the queries and keys occupy different grids. For example, The video diffusion latents $\boldsymbol{z}_v$, compressed by the 3D VAE, live on a (T_v^l, H_v^l, W_v^l) grid, while $\boldsymbol{H}_v^{dit}$ are arranged on a (N, h_s, w_s) semantic grid ($n_v = h_s \times w_s$ spatial tokens per keyframe). The two grids differ along all three axes, so tokens representing the same space-time point get different indices under standard RoPE, leading to misaligned encoding and incorrect attention.

Thus, we introduce the Relative Semantic RoPE, which rescales the position indices of both grids into a shared coordinate system. Regarding $\texttt{VCAttn}(\cdot)$, given a latent query at grid index (i_t, i_h, i_w) and a planned semantic key at (j_t, j_h, j_w), the query and key position indices are $\boldsymbol{p}_v^q = (i_t,\ i_h,\ i_w)$ and $\boldsymbol{p}_v^k = \left(j_t \cdot \frac{T_v^l}{N},\ \ j_h \cdot \frac{H_v^l}{h_s},\ \ j_w \cdot \frac{W_v^l}{w_s}\right)$. We then inject $\boldsymbol{H}_v^{dit}$ to the DiT:

$$\begin{aligned}
\texttt{RoPE}(\boldsymbol{h}_*, p) &= \boldsymbol{h}_* \odot \cos(p \cdot \boldsymbol{\omega}) + \texttt{Rotate}(\boldsymbol{h}_*) \odot \sin(p \cdot \boldsymbol{\omega}), \\
\texttt{RoPE}_{3D}(\boldsymbol{h}, \boldsymbol{p}) &= \big[\texttt{RoPE}(\boldsymbol{h}^{(t)}, p_t);\ \ \texttt{RoPE}(\boldsymbol{h}^{(h)}, p_h);\ \ \texttt{RoPE}(\boldsymbol{h}^{(w)}, p_w)\big], \\
\texttt{VCAttn}(\boldsymbol{z}_v, \boldsymbol{H}_v^{dit}) &= \texttt{Attn}(\texttt{RoPE}_{3D}(\boldsymbol{z}_v, \boldsymbol{p}_v^q), \texttt{RoPE}_{3D}(\boldsymbol{H}_v^{dit}, \boldsymbol{p}_v^k), \boldsymbol{H}_v^{dit}),
\end{aligned} \tag{6}$$

where $\texttt{Attn}(\cdot)$ is the attention operation. $\texttt{RoPE}_{3D}(\cdot)$ is applied per attention head. For each head, the query or key vector $\boldsymbol{h} \in \mathbb{R}^{d_h}$ is partitioned into $[\boldsymbol{h}^{(t)}; \boldsymbol{h}^{(h)}; \boldsymbol{h}^{(w)}]$ corresponding to the temporal, height, and width axes. The frequency vector $\omega_k = \frac{1}{\theta^{2k/d_h}}$ (θ is a hyperparameter) and $\texttt{Rotate}(\cdot)$ swaps adjacent pairs with sign flips following the standard RoPE [68]. As our RoPE encodes only relative position, it endows the cross-attention with a soft spatio-temporal locality prior that preserves the keyframe-level structure of the blueprints, rather than flattening it into a global context.

Symmetrically, $\texttt{ACAttn}(\cdot)$ uses $\boldsymbol{z}_a \in \mathbb{R}^{T_a^l \times d}$ as queries and $\boldsymbol{H}_a^{dit}$ as keys and values. Since audio carries no spatial structure, the alignment reduces to 1D RoPE:

$$\texttt{ACAttn}(\boldsymbol{z}_a, \boldsymbol{H}_a^{dit}) = \texttt{Attn}(\texttt{RoPE}(\boldsymbol{z}_a, p_a^q), \texttt{RoPE}(\boldsymbol{H}_a^{dit}, p_a^k), \boldsymbol{H}_a^{dit}), \tag{7}$$

where $p_a^q(i) = i$ for the i-th audio latent and $p_a^k(k) = k \cdot T_a^l / L_a$ for the k-th planned audio token, aligning both onto the same temporal axis. Notably, previous cross-modal RoPE strategies in joint video-audio generation [48, 70, 100, 28] only handle 1D temporal scaling between homogeneous video and audio latent streams that share the same backbone architecture. Our RoPE goes further by operating in full 3D, simultaneously resolving temporal and spatial mismatches between two fundamentally heterogeneous features (planned semantic tokens and diffusion latents).

3.3 Training

Baton is trained in three stages: (1) **Stage 1 (VA-Planner Pretraining).** The VA-Planner aims to autoregressively reason over user prompts and predict planned tokens that capture the perceptual structure of the target content. We initialize the MLLM from Qwen3 [95] and train the entire VA-Planner ($\texttt{MLLM+SMLP}_{v/a}(\cdot)$). Given ground-truth video and audio, we extract target continuous features $\boldsymbol{F}_v^{gt} \in \mathbb{R}^{L_v \times D_s}$ and $\boldsymbol{F}_a^{gt} \in \mathbb{R}^{L_a \times D_a}$ from the penultimate layers [49] of frozen SigLip2 and WavTokenizer. The VA-Planner is supervised as follows:

$$\mathcal{L}_{plan} = \sum_{t=1}^{N} \sum_{i=1}^{n_v} \|\boldsymbol{H}_{v,(t,i)}^{sem} - \boldsymbol{F}_{v,(t,i)}^{gt}\|_2^2 + \sum_{m=1}^{M} \sum_{j=1}^{n_a} \|\boldsymbol{H}_{a,(m,j)}^{sem} - \boldsymbol{F}_{a,(m,j)}^{gt}\|_2^2, \tag{8}$$

where $\boldsymbol{H}_{v,(t,i)}^{sem}$ is the planned video token at the i-th spatial position of the t-th keyframe and $\boldsymbol{H}_{a,(m,j)}^{sem}$ is the planned audio token at the j-th position of the m-th audio chunk. By contrast to planning over

Table 1: Quantitative comparisons with previous open-source methods on Verse-Bench and Sem100. In the table elements a/b, a and b refer to the results on Verse-Bench and Sem100.

Model	AQ↑	IQ↑	DD↑	ID↑	PQ↑	CU↑	M-WER↓	Sync-C↑	Sync-D↓	DeSync↓	P-Acc↑
Ovi [48]	0.56/0.42	0.67/0.65	0.49/0.36	0.90/0.88	6.30/6.74	5.91/6.23	0.75/0.66	3.34/5.07	8.78/8.25	0.54/1.06	0.74/0.46
JavisGPT [46]	0.34/0.28	0.45/0.38	0.30/0.32	0.32/0.45	5.37/5.12	4.77/5.06	4.76/3.96	0.65/0.69	11.27/12.34	1.16/1.18	0.43/0.25
UniAVGen [100]	0.55/0.31	0.67/0.58	0.47/0.32	0.86/0.82	4.94/6.68	6.11/5.83	1.05/0.72	1.95/4.80	9.57/9.83	0.87/1.14	0.76/0.44
LTX-2 [21]	0.53/0.48	0.66/0.65	**0.71**/0.38	0.92/0.91	6.54/7.14	6.01/6.91	0.64/0.58	3.77/6.26	8.16/7.72	**0.27**/0.97	0.85/0.62
MOVA [70]	0.54/0.44	0.64/0.68	0.56/0.40	0.90/0.87	6.71/6.95	6.19/6.94	1.49/0.86	3.52/5.78	10.58/8.90	0.64/1.05	0.79/0.55
Ours	**0.58/0.54**	**0.68/0.73**	0.64/**0.48**	**0.93/0.94**	**6.79/7.53**	**6.24/7.16**	**0.18/0.14**	**4.26/8.14**	**7.68/6.85**	0.57/**0.68**	**0.88/0.82**

discrete tokens, regressing continuous features preserves richer semantic structure. (2) **Stage 2 (DiT Adaptation).** To allow the DiT to learn the semantic feature distribution without being confounded by planner prediction noise, we initialize the DiT from Ovi [48] and feed the ground-truth $\boldsymbol{F}_v^{gt}$, $\boldsymbol{F}_a^{gt}$ (projected via Latent-MLPs) directly into $\texttt{VCAttn}(\cdot)$ and $\texttt{ACAttn}(\cdot)$. We use the flow matching loss [45] to train the DiT $\hat{\boldsymbol{v}}_\theta^{v/a}(\cdot)$ (including $\texttt{LMLP}_{v/a}(\cdot)$):

$$\mathcal{L}_{FM} = \mathbb{E}_{\boldsymbol{z}_0^{v/a}, \boldsymbol{z}_1^{v/a}, t}\big[\|\hat{\boldsymbol{v}}_\theta^v(\boldsymbol{z}_t^v, \boldsymbol{z}_t^a, t, \boldsymbol{c}, \boldsymbol{F}_v^{gt}) - (\boldsymbol{z}_1^v - \boldsymbol{z}_0^v)\|_2^2 + \|\hat{\boldsymbol{v}}_\theta^a(\boldsymbol{z}_t^v, \boldsymbol{z}_t^a, t, \boldsymbol{c}, \boldsymbol{F}_a^{gt}) - (\boldsymbol{z}_1^a - \boldsymbol{z}_0^a)\|_2^2\big], \quad (9)$$

where $\boldsymbol{z}_1^{v/a} \sim \mathcal{N}(0, I)$. $\boldsymbol{z}_0^{v/a}$ and $\boldsymbol{c}$ are the VAE-encoded clean latents and the conditioning signals. (3) **Stage 3 (Joint Fine-tuning).** The VA-Planner and DiT are connected. The VA-Planner is frozen, and the DiT is trainable. The DiT now receives $\boldsymbol{H}_v^{sem}$ and $\boldsymbol{H}_a^{sem}$ as conditioning, and training continues with Eq.9. This stage bridges the gap between the clean encoder features in Stage 2 and the imperfect planner predictions, mitigating exposure bias and ensuring robust generation.

4 Experiments

4.1 Implementation Details

Our training dataset (1.5 million video-audio clips) is aggregated from OpenHuman-Vid [42], AudioCaps [36], WavCaps [51], and videos collected from the internet (Appx. A.5). Following previous works [70], we evaluate our model on Verse-Bench [85]. We conduct additional experiments on 100 unseen videos (10 seconds long) with more complex prompts, referred to the Sem100, selected from the internet to assess the semantic reasoning capability of our model. Our DiT and VA-Planner are initialized by Ovi [48] and Qwen3-8B [95]. Our model is trained for 10 epochs in 3 training stages on 160 NVIDIA H800 GPUs, with a batch size of 1 per GPU. The learning rate is $1e-5$ (Appx. A.6).

4.2 Comparison with State-of-the-Art Methods

Quantitative results. Following previous works [28, 100], we utilize AQ [28], IQ [32], DD [32], and ID [28] to assess the video quality. We further apply PQ [71], CU [71], M-WER (Multi-Speaker Word Error Rate) [58], Sync-C [11], Sync-D [11], and DeSync [33] to assess the audio quality and video-audio synchronization. P-Acc evaluates the prompt following accuracy (Appx. A.2). We perform quantitative comparisons with previous methods [46, 48, 21, 70, 100] on Verse-Bench [85] and Sem100, as shown in Table 1. Compared to the leading competitor LTX-2 [21], Baton achieves comparable results on Verse-Bench, where the text prompts mostly describe simple, single-event scenarios that do not require deep semantic reasoning. The advantage becomes prominent on Sem100, whose prompts involve complex sequential events, intricate human-object interactions, and multi-speaker dialogues that demand strong semantic reasoning. On Sem100, Baton outperforms LTX-2 by 32% in P-Acc, 76% in M-WER, and 30% in DeSync. The M-WER gap is particularly striking, as multi-speaker scenarios require the model to reason about which character speaks what content and when, exactly the kind of localized, temporally grounded semantics that our planned tokens provide, but global text embeddings cannot decompose. The large P-Acc and M-WER gaps confirm that explicit semantic planning is essential for complex prompts where global text embeddings fail to decompose multi-stage actions and multi-speaker dialogues into temporally grounded guidance.

Qualitative Results. The qualitative results are shown in Fig. 3. Notably, since UniAVGen and MOVA only open-source I2VA, we use the first frame of videos generated by Baton as their reference images. All competitors fail to handle prompts involving sequential events or human-object interactions, indicating that global text embeddings lack structured cross-modal semantic comprehension. Ovi [48] and UniAVGen [100] suffer from severe body distortion, a result of



Figure 3: Qualitative comparisons with previous open-source methods. Please refer to the demo video for audio. More results are in the Appx. A.7.

the two denoising trajectories drifting into adversarial dynamics under complex prompts, where the video and audio branches independently interpret the vague global signal and pull the joint distribution in conflicting directions, ultimately distorting the video. LTX-2 [21] and MOVA [70] exhibit poor subject consistency and audio-video desynchronization. For instance, the basketball in the boy's hands changes color or disappears. The global text embeddings cannot model how each object and action should coherently manifest and temporally progress across both modalities, leaving the two trajectories uncoordinated over long horizons. In contrast, since Baton establishes cross-modal semantic coordination before denoising, both generation trajectories are anchored to a shared, pre-aligned semantic roadmap, preventing adversarial drift and yielding stable audio-video contents.

The essential distinction among these benchmarks lies in the complexity of prompts. Verse-Bench prompts are basically single-sentence, single-event descriptions that describe one atomic action with one sound source. For such prompts, off-the-shelf text encoders (T5, Qwen) already compress the full semantic content into a global embedding with minimal loss, because there is no multi-step temporal structure to decompose and no cross-modal causal dependency to reason about. All models thus perform comparably on Verse-Bench, and Baton's planning mechanism yields only marginal gains because the failure mode it addresses does not manifest. Sem100 prompts are more complex, involving the most demanding multi-stage compositional prompts requiring causal reasoning over complex action sequences and multi-speaker dialogues with speaker-specific content at specific temporal positions. Baton's advantage consistently grows with prompt complexity, validating that explicit semantic planning becomes essential once prompts exceed the capacity of global text embeddings.

Regarding the DeSync in Table 1, Baton is marginally below its own DiT backbone Ovi on Verse-Bench, and LTX-2 achieves the best DeSync. Baton's slight regression relative to Ovi is because Stage-2/3 fine-tuning introduces VCAttn/ACAttn that redistributes the DiT's attention toward planned tokens. For single-event prompts where text embeddings already provide complete guidance, this redistribution dilutes the original alignment signal. LTX-2's strong Verse-Bench DeSync stems from its 19B DiT capacity and larger training data, which enable precise memorization of simple single-event audio-visual patterns. However, this advantage collapses on complex prompts in Sem100, revealing that scaling DiT alone cannot handle multi-event synchronization. By contrast, Baton trades only a marginal DeSync increase on Verse-Bench for substantial gains on complex prompts.

Comparison with Commercial Models. We compare our model with commercial models [7, 18, 40, 1], as shown in Appx. A.8. We can see that although commercial models outperform Baton in video fidelity and audio aesthetics, they show comparable prompt-following capabilities, showing the superiority of Baton in handling scenarios demanding semantic reasoning.

4.3 Ablation Study

VA-Planner. We conduct an ablation study to validate the contributions of VA-Planner in Baton, as shown in Fig.5 (a) and Table 2. Notably, all quantitative ablation studies are on the Sem100

Table 2: Ablation study on VA-Planner on Sem100. Only Planned Tokens removes global text embeddings and only injects planned tokens into the DiT. *w/* Prompt Enhancement replaces VA-Planner with Qwen3-refined prompts. *w/* Frozen LLM feeds frozen Qwen3-8B hidden states into DiT without VA-Planner. Tower refers to the dual semantic alignment towers.

Model	AQ↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑
w/o VA-Planner (Ovi [48])	0.44	6.88	6.45	0.62	0.98	0.51
Only Planned Tokens	0.53	7.38	7.04	0.17	0.73	0.78
w/ Prompt Enhancement (PE-Qwen3-235B-A22B [95])	0.51	6.97	6.64	0.56	0.93	0.62
w/ Frozen LLM (Qwen3-8B [95])	0.52	7.27	6.68	0.48	0.86	0.67
w/o Learnable Query	0.52	7.25	6.88	0.27	0.74	0.79
w/o Tower	0.48	6.92	6.53	0.45	0.78	0.69
w/o RoPE in Tower	0.42	6.57	6.24	0.68	1.05	0.44
w/ TA-Tok [22]+WavTokenizer	0.47	7.16	6.85	0.32	0.81	0.68
w/ DINOv3 [64]	0.50	7.36	7.02	0.29	0.74	0.77
w/ Beats [10]	0.48	7.28	6.77	0.36	0.78	0.73
w/ Unified Tokens [46]	0.44	7.12	6.73	0.41	0.85	0.65
Ours	**0.54**	**7.53**	**7.16**	**0.14**	**0.68**	**0.82**

dataset, and all ablated models are trained on the same dataset under identical settings. *w/* TA-Tok+WavTokenizer replaces the continuous SigLip2 and WavTokenizer targets with their discrete counterparts (TA-Tok [22] and discrete WavTokenizer). *w/* DINOv3 replaces SigLip2 with DINOv3 [64] as the video alignment target. *w/* Beats replaces WavTokenizer with Beats [10] as the audio alignment target. *w/* Unified Tokens [46] replaces separate video and audio planning with a single unified token sequence that jointly describes both modalities. By analyzing the results, we can obtain the following observations: (1) Removing VA-Planner significantly degrades performance, particularly in M-WER, DeSync, and P-Acc, indicating that VA-Planner can significantly improve video-audio synchronization and prompt following ability by providing dedicated planned tokens. In contrast, *w/* Prompt Enhancement and *w/* Frozen LLM still suffer from accurate and stable video-audio generation due to their coarse guidance. Notably, Only Planned Tokens already achieves strong results, confirming they are the primary semantic driver. The gap to the full model shows that global text embeddings and planned tokens are complementary. Global embeddings provide a holistic prior for coherence, while planned tokens supply fine-grained, position-specific structure. (2) Removing RoPE in Tower causes the most severe degradation, because the towers without temporal positional alignment produce temporally misaligned planned tokens that mislead denoising. *w/o* Tower confirms that the video plan without bidirectional cross-modal attention is entirely blind to audio context (due to causal masking in the MLLM), producing two independently derived plans that lack mutual consistency and inevitably diverge during denoising. *w/o* Learnable Query shows that query-based distillation can gain more semantic details. (3) Replacing continuous alignment targets with discrete TA-Tok [22]+WavTokenizer leads to clear drops across most metrics, because discrete quantization inevitably loses the fine-grained perceptual nuances that continuous features preserve, reducing the semantic richness available to guide each denoising step. *w/* Unified Tokens [46] yields even larger degradation. Since video and audio possess fundamentally different spatio-temporal structures, collapsing them into a single token sequence forces both modalities to share one bottleneck, blurring modality-specific semantics and undermining the fine-grained cross-modal coordination that separate planning naturally provides. (4) Both DINOv3 [64] and Beats [10] underperform SigLip2 and WavTokenizer. It indicates that SigLip2's text-aligned visual features offer stronger semantic grounding than DINOv3's self-supervised features. While Beats is trained for audio classification and encodes category-level abstractions, WavTokenizer's reconstruction-oriented features preserve the fine-grained acoustic details that the diffusion model requires for audio synthesis.

We further ablate the injection of planned tokens, keyframe FPS, and the orders of T_v^{tag} and T_a^{tag}, as shown in Appx. A.9. Three key findings emerge. (1) Our injection outperforms other alternatives, because a coarse-to-fine hierarchy lets the text prior and fine-grained planned cues complement rather than dilute each other. (2) Keyframe FPS exhibits semantic saturation at FPS=6, as the keyframe-level blueprints already sufficiently cover the temporal semantics for denoising. (3) Placing T_v^{tag} before T_a^{tag} exploits the MLLM's causal attention so that audio plan conditions on the video plan, matching the natural audio-follows-video dependency. Details are in Appx. A.9.

RS-RoPE. We conduct an ablation study on RS-RoPE, as shown in Table 3 and Fig. 5 (b). *w/o* RS-RoPE causes drastic degradation, even worse than *w/o* VA-Planner. The reason is that the cross-

Table 3: Ablation study on RS-RoPE and different backbones. Temporal RoPE [70] only scales the position indices in the temporal axis. Baton-Qwen3-X uses Qwen3-X to initialize the MLLM. † denotes that single-GPU (80GB) inference is infeasible and 2-GPU FSDP parallelism is required.

Category	Model	AQ↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑	Per-GPU Mem↓
RS-RoPE	*w/o* RS-RoPE	0.40	6.42	6.15	0.65	1.08	0.46	68.4G
	w/ Temporal RoPE [70]	0.51	7.39	6.92	0.41	0.75	0.73	68.4G
MLLM Backbones	Baton-Qwen3-VL-8B-Instruct [2]	0.44	6.90	6.51	0.45	0.84	0.72	71.2G
	Baton-Qwen3-Omni-30B-A3B [93]	0.41	6.76	6.28	0.65	1.02	0.65	74.6G
	Baton-Qwen3-4B [95]	0.46	7.32	6.90	0.31	0.82	0.68	63.5G
	Baton-Qwen3-30B-A3B-Instruct [95]	0.55	7.51	7.12	0.16	0.61	0.82	71.8GG
	Baton-Qwen3-32B [95]	**0.58**	**7.86**	**7.29**	**0.13**	**0.54**	**0.85**	2×69.4G†
DiT Backbones	LTX-2 [21]	0.48	7.14	6.91	0.58	0.97	0.62	48.5G
	LTX-2 [21]+Baton-Qwen3-8B (*w/o* DiT)	0.56	7.49	7.21	0.16	0.64	0.83	64.6G
Final	Baton-Qwen3-8B [95]	0.54	7.53	7.16	0.14	0.68	0.82	68.4G

attention degenerates into purely content-based matching that treats all planned tokens as an orderless bag, collapsing the structured keyframe-level plan into a vague global signal that misleads denoising. *w/* Temporal RoPE [70] recovers much of the performance but still falls short, as it only aligns the temporal axis while ignoring the spatial mismatch between the semantic grid and the latent grid, preventing latents from attending to their spatially corresponding planning cues.

Different Backbones. We ablate the backbones of the VA-Planner and DiT, as shown in Fig. 5 (c) and Table 3. We have the following observations: (1) Qwen3-VL/Omni initialization are worse than ours, despite their larger pretraining scope. Qwen3-VL is pretrained for visual understanding with features heavily shaped toward high-level recognition rather than the spatial-perceptual structure required by SigLip2 alignment. Migrating such a distribution toward SigLip2's spatial domain is harder than training a general-purpose Qwen3 from scratch, as the pretrained weights resist the new alignment objective. Qwen3-Omni suffers similarly because its audio capabilities are pretrained on speech-centric data with a fixed voice identity, yielding an output distribution that lacks acoustic diversity. Forcing this narrow distribution to cover the broad range of environmental sounds and diverse timbres demands more data to overcome the pretrained bias. (2) Scaling the MLLM backbone from 4B to 32B improves all metrics, confirming that stronger reasoning capacity leads to higher-quality planned tokens. However, Qwen3-32B incurs significant training and inference overhead. Qwen3-30B-A3B performs similarly to the dense Qwen3-8B despite more parameters, as MoE's sparse routing activates different experts per token, fragmenting features and weakening the holistic cross-token reasoning needed for globally consistent long-horizon planning. Ours strikes a practical balance between performance and computational cost. (3) Replacing the DiT backbone from Ovi [48] to LTX-2 [21] validates the robustness of our model across different DiT backbones.

Training Stages. We compare our model in different training stages, as shown in Table 9. We can see that skipping Stage-2 causes significant degradation. The root reason is that the DiT's cross-attention layers have never been exposed to semantic features. Directly receiving imperfect planner predictions in Stage-3 introduces a distribution shock that the model struggles to absorb. Our three-stage curriculum avoids this by letting Stage-2 first teach the DiT to leverage clean ground-truth features, establishing a stable semantic interface, then Stage-3 progressively adapts it to VA-Planner predictions, effectively reducing exposure bias. Furthermore, Stage-2 outperforms our full model because it conditions on ground-truth SigLip2 and WavTokenizer features, which provide perfect semantic blueprints free from any prediction noise. However, ground-truth perceptual features are unavailable at inference since they require access to the target video and audio. Stage-3 closes this gap by adapting the DiT to realistic VA-Planner predictions, achieving robust generation under practical conditions. More ablation studies are in Appx. A.9.

Planned Token Diversity. The VA-Planner is indeed deterministic for a given prompt, but the DiT's denoising process starts from different random noise across seeds. Planned tokens encode keyframe-level semantic content (SigLip2 visual semantics, WavTokenizer acoustic structure) that specifies what happens when, without constraining pixel-level realization. The DiT retains full stochastic freedom over textures and fine-grained audio timbres. We conduct a quantitative diversity evaluation. We evaluate diversity using the same ablation variants from Table 2. In particular, we generate 5 outputs per prompt with different random seeds on two prompt categories. "Simple" denotes single-sentence, single-event descriptions. "Complex" denotes multi-event prompts involving sequential events, complex actions, or multi-person interaction with environmental elements. We

Table 4: Intra-prompt generation diversity (5 seeds $\times$ 30 prompts per setting).

Setting	Model	Intra-LPIPS↑	DINO-Div↑
Simple	*w/o* VA-Planner (Ovi)	0.26	0.38
Simple	*w/* Frozen LLM	0.25	0.36
Simple	Only Planned Tokens	0.22	0.33
Simple	Baton (Ours)	0.24	0.34
Complex	*w/o* VA-Planner (Ovi)	0.38	0.49
Complex	*w/* Frozen LLM	0.33	0.46
Complex	Only Planned Tokens	0.20	0.25
Complex	Baton (Ours)	0.23	0.27

measure two intra-prompt diversity metrics. Intra-LPIPS computes the mean pairwise LPIPS distance across all pairs of generated videos per prompt (frame-level LPIPS averaged temporally), capturing perceptual diversity in visual appearance. DINO-Div extracts DINOv2 features per frame, averages them temporally into a video-level representation, and computes the mean pairwise cosine distance across 5 outputs, capturing higher-level semantic diversity. Table 4 reveals a favorable trade-off between diversity and semantic accuracy. On simple prompts, Baton's DINO-Div remains close to *w/o* VA-Planner, while Intra-LPIPS stays high, indicating that planned tokens anchor the scene layout without suppressing low-level variation. On complex prompts, Baton's DINO-Div drops more noticeably. However, the baselines' inflated DINO-Div does not reflect genuine creative diversity. The global text embedding is too lossy to fully specify temporal event ordering, so different noise seeds lead to semantically divergent and often incorrect outputs. In contrast, Baton's planned tokens provide position-specific conditioning at each keyframe and audio chunk, collapsing the output distribution to the correct semantic mode. Meanwhile, Intra-LPIPS remains nearly stable, confirming that the DiT retains full freedom over pixel-level details, all governed by the initial noise.

4.4 Applications and User Study

Speed and GPU Resource. We compare the inference latency and GPU cost between Baton and previous models, as shown in Appx.A.10. Compared to our DiT backbone Ovi [48], Baton introduces 28% additional GPU memory, while achieving over 36% improvement in DeSync and 78% in P-Acc. It indicates that the planning overhead is a highly favorable trade-off for the substantial quality gains, highlighting its superiority in handling semantically complex scenarios.

Multi-Speakers. We experiment on multi-speaker scenarios, as shown in Appx.A.12. We can see that Baton is capable of joint video-audio contents involving multi-character interaction.

Complex Scene. We experiment in semantically complex scenarios, as shown in Appx. A.14. Baton can handle long-horizon scenes involving multiple sequential actions and complex interactions.

User Study. We conduct a user study on 30 selected videos. The participants are university students and faculty. In each case, participants are first presented with the video/audio prompt. We then provide two videos, one of which is generated by Baton and the other is synthesized by a competitor. Participants are asked to answer questions: V-A/A-A: "Which one has better video/audio alignment with the prompts?" I-C/B-C/V-A-S: "Which one has better ID/background consistency/video-audio synchronization?" Appx.A.11 shows the superiority of our model in subjective evaluation.

5 Conclusion

In this paper, we proposed Baton, a video diffusion transformer with a dedicated MLLM to jointly synthesize stable video-audio, even for scenarios requiring complex semantic reasoning. To explicitly disentangle semantic reasoning from synthesis before denoising, Baton first introduced the VA-Planner, an MLLM equipped with dual semantic alignment towers, which reasoned over user prompts to produce a pair of mutually aligned video and audio planned tokens as keyframe-level blueprints, anchoring both denoising trajectories to a shared semantic roadmap. To address the spatio-temporal mismatch between planned tokens and diffusion latents, Baton further introduced Relative Semantic RoPE. Extensive experiments demonstrated that Baton significantly outperforms open-source methods, particularly on complex prompts demanding semantic reasoning, and achieves comparable prompt-following capability to commercial models.

6 Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62472098).

References

[1] Alibaba Group. Wan 2.7. `https://wan.video/`, 2026.

[2] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, S. Liu, Y. Wang, and J. Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

[5] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[6] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024.

[7] ByteDance. Seedance 2.0. `https://seed.bytedance.com/en/seedance2_0`, 2026.

[8] J. Chen, Z. Xu, X. Pan, Y. Hu, C. Qin, T. Goldstein, L. Huang, T. Zhou, S. Xie, S. Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[9] J. Chen, L. Xue, Z. Xu, X. Pan, S. Yang, C. Qin, A. Yan, H. Zhou, Z. Chen, L. Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025.

[10] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

[11] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016.

[12] J. Cui, H. Li, Y. Zhan, H. Shang, K. Cheng, Y. Ma, S. Mu, H. Zhou, J. Wang, and S. Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *CVPR*, 2025.

[13] C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[14] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[15] discus0434. Aesthetic predictor v2.5. `https://github.com/discus0434/aesthetic-predictor-v2-5`, 2024.

[16] Q. Gan, R. Yang, J. Zhu, S. Xue, and S. Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.

[17] Y. Ge, S. Zhao, J. Zhu, Y. Ge, K. Yi, L. Song, C. Li, X. Ding, and Y. Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[18] Google DeepMind. Veo 3. `https://deepmind.google/models/veo/`, 2025.

[19] Google DeepMind. Gemini 3.1. `https://deepmind.google/models/gemini/`, 2026.

[20] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

[21] Y. HaCohen, B. Brazowski, N. Chiprut, Y. Bitterman, A. Kvochko, A. Berkowitz, D. Shalem, D. Lifschitz, D. Moshe, E. Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

[22] J. Han, H. Chen, Y. Zhao, H. Wang, Q. Zhao, Z. Yang, H. He, X. Yue, and L. Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025.

[23] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[24] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[25] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022.

[26] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[27] J. Hu, S. Zhao, Q.-G. Chen, X. Qiu, J. Liu, Z. Xu, W. Luo, K. Zhang, and Y. Lu. Omni-view: Unlocking how generation facilitates understanding in unified 3d model based on multiview images. In *ICLR*, 2026.

[28] T. Hu, Z. Yu, G. Zhang, Z. Su, Z. Zhou, Y. Zhang, Y. Zhou, Q. Lu, and R. Yi. Harmony: Harmonizing audio and video generation through cross-task synergy. *arXiv preprint arXiv:2511.21579*, 2025.

[29] L. Huang, Y. Xie, H. Xu, T. Gu, C. Zhang, G. Song, Z. Li, X. Zhao, L. Luo, and G. Sapiro. Plan-x: Instruct video generation via semantic planning. *arXiv preprint arXiv:2511.17986*, 2025.

[30] Y. Huang, S. Tu, X. Yan, Z. Yang, J. Han, H. Xu, K. Pan, Y.-G. Jiang, and Z. Wu. Va-judger: Reward modeling from human preference feedback for joint video-audio generation. *arXiv preprint arXiv:2608.18607*, 2026.

[31] Y. Huang, S. Tu, X. Yan, J. Zhan, Z. Yang, Z. Xing, H. Zhang, T. Zhang, Y.-G. Jiang, and Z. Wu. Agentic visual generation: From generative models to agentic control. *arXiv preprint arXiv:2609.06758*, 2026.

[32] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

[33] V. Iashin, W. Xie, E. Rahtu, and A. Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP*, 2024.

[34] S. Ji, Z. Jiang, W. Wang, Y. Chen, M. Fang, J. Zuo, Q. Yang, X. Cheng, Z. Wang, R. Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

[35] X. Ji, X. Hu, Z. Xu, J. Zhu, C. Lin, Q. He, J. Zhang, D. Luo, Y. Chen, Q. Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In *CVPR*, 2025.

[36] C. D. Kim, B. Kim, H. Lee, and G. Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[37] D. Kim, J. He, Q. Yu, C. Yang, X. Shen, S. Kwak, and L.-C. Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In *ICCV*, 2025.

[38] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[39] Z. Kong, F. Gao, Y. Zhang, Z. Kang, X. Wei, X. Cai, G. Chen, and W. Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025.

[40] Kuaishou Technology. Kling 3.0. `https://kling.ai/app`, 2026.

[41] J. Leng, S. Tu, H. Cao, S. Xie, D. Dong, Z. Wu, and Y.-G. Jiang. Preference score distillation: Leveraging 2d rewards to align text-to-3d generation with human preference. *arXiv preprint arXiv:2603.01594*, 2026.

[42] H. Li, M. Xu, Y. Zhan, S. Mu, J. Li, K. Cheng, Y. Chen, T. Chen, M. Ye, J. Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *CVPR*, 2025.

[43] C. Liao, L. Liu, X. Wang, Z. Luo, X. Zhang, W. Zhao, J. Wu, L. Li, Z. Tian, and W. Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

[44] H. Lin, J. Cho, A. Zadeh, C. Li, and M. Bansal. Bifrost-1: Bridging multimodal llms and diffusion models with patch-level clip latents. *arXiv preprint arXiv:2508.05954*, 2025.

[45] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[46] K. Liu, J. Li, Y. Sun, S. Wu, J. Gao, D. Zhang, W. Zhang, S. Jin, S. Yu, G. Zhan, et al. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. *arXiv preprint arXiv:2512.22905*, 2025.

[47] K. Liu, W. Li, L. Chen, S. Wu, Y. Zheng, J. Ji, F. Zhou, J. Luo, Z. Liu, H. Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025.

[48] C. Low, W. Wang, and C. Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

[49] B. Ma, Z. Zong, G. Song, H. Li, and Y. Liu. Exploring the role of large language models in prompt encoding for diffusion models. *arXiv preprint arXiv:2406.11831*, 2024.

[50] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, X. Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, 2025.

[51] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[52] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021.

[53] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

[54] X. Pan, S. N. Shukla, A. Singh, Z. Zhao, S. K. Mishra, J. Wang, Z. Xu, J. Chen, K. Li, F. Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

[55] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[56] X. Qiu, Y. Cui, G. Zhang, J. Li, J. Hu, X. Zhang, Y. Li, S. Liu, M. Yang, Y. Shi, et al. Hydra: Unifying multi-modal generation and understanding via representation-harmonized tokenization. *arXiv preprint arXiv:2603.15228*, 2026.

[57] L. Qu, H. Zhang, Y. Liu, X. Wang, Y. Jiang, Y. Gao, H. Ye, D. K. Du, Z. Yuan, and X. Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *CVPR*, 2025.

[58] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

[59] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[60] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[61] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023.

[62] J. Shi, Q. Hou, J. Leng, J. Wang, Y. Zhang, H. Lu, Z. Ye, and Z. Wu. Eva-cap: Optimizing audiovisual video captioning via event-centric alignment. *arXiv preprint arXiv:2610.22186*, 2026.

[63] J. Shi, J. Wang, Z. You, B. He, and Z. Wu. Videoloom: A video large language model for joint spatial-temporal understanding. *arXiv preprint arXiv:2601.07290*, 2026.

[64] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[65] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[66] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[67] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[68] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

[69] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Y. Wang, Y. Rao, J. Liu, T. Huang, and X. Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024.

[70] O. Team, D. Yu, M. Chen, Q. Chen, Q. Luo, Q. Wu, Q. Cheng, R. Li, T. Liang, W. Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

[71] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

[72] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[73] S. Tu, Q. Dai, Z.-Q. Cheng, H. Hu, X. Han, Z. Wu, and Y.-G. Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, 2024.

[74] S. Tu, Q. Dai, Z. Wu, Z.-Q. Cheng, H. Hu, and Y.-G. Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *ICCV*, 2023.

[75] S. Tu, Q. Dai, Z. Zhang, S. Xie, Z.-Q. Cheng, C. Luo, X. Han, Z. Wu, and Y.-G. Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024.

[76] S. Tu, T. Guan, and L. Kuang. Multiple biological granularities network for person re-identification. In *ICMR*, 2022.

[77] S. Tu, Y. Pan, Y. Huang, X. Han, Z. Xing, Q. Dai, C. Luo, Z. Wu, and Y.-G. Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

[78] S. Tu, Y. Pan, Y. Huang, X. Han, Z. Xing, Q. Dai, K. Qiu, C. Luo, and Z. Wu. Flashportrait: 6x faster infinite portrait animation with adaptive latent prediction. *arXiv preprint arXiv:2512.16900*, 2025.

[79] S. Tu, Q. Tian, Y. Huang, Y. Wu, X. Han, K. Pan, W. Kong, J. Xiong, J.-W. Zhang, Z. Wu, et al. Prism: Dynamic sparse attention for native 2k joint video-audio generation model training. *arXiv preprint arXiv:2610.05416*, 2026.

[80] S. Tu, Z. Xing, X. Han, Z.-Q. Cheng, Q. Dai, C. Luo, and Z. Wu. Stableanimator: High-quality identity-preserving human image animation. In *CVPR*, 2025.

[81] S. Tu, Z. Xing, X. Han, Z.-Q. Cheng, Q. Dai, C. Luo, Z. Wu, and Y.-G. Jiang. Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. *arXiv preprint arXiv:2507.15064*, 2025.

[82] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

[83] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[84] B. Wang, Z. Yue, F. Zhang, S. Chen, L. Bi, J. Zhang, X. Song, K. Y. Chan, J. Pan, W. Wu, et al. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning. *arXiv preprint arXiv:2505.07538*, 2025.

[85] D. Wang, W. Zuo, A. Li, L.-H. Chen, X. Liao, D. Zhou, Z. Yin, X. Dai, D. Jiang, and G. Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

[86] W. Wang, J. Liu, Z. Lin, J. Yan, S. Chen, C. Low, T. Hoang, J. Wu, J. H. Liew, H. Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024.

[87] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[88] X. Wang, S. Zhou, K. Wang, L. Xu, X. Qiu, T. He, and M. Li. Why do dit editors drift? plug-and-play low frequency alignment in vae latent space. *arXiv preprint arXiv:2605.08250*, 2026.

[89] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023.

[90] Y. Wu, Z. Zhang, J. Chen, H. Tang, D. Li, Y. Fang, L. Zhu, E. Xie, H. Yin, L. Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[91] J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z. Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[92] Y. Xie, T. Gu, Z. Li, C. Zhang, G. Song, X. Zhao, C. Liang, J. Jiang, H. Xu, and L. Luo. X-streamer: Unified human world modeling with audiovisual interaction. *arXiv preprint arXiv:2509.21574*, 2025.

[93] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[94] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, Y. Yao, and S. Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

[95] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[96] S. Yang, Z. Kong, F. Gao, M. Cheng, X. Liu, Y. Zhang, Z. Kang, W. Luo, X. Cai, R. He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

[97] Z. Yang, S. Tu, L. Zhang, Q. Dai, Y.-G. Jiang, and Z. Wu. Arcflow: Unleashing 2-step text-to-image generation via high-precision non-linear flow distillation. *arXiv preprint arXiv:2602.09014*, 2026.

[98] J. Yao, B. Yang, and X. Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025.

[99] G. Zhang, X. Qiu, Y. Cui, T. Song, C. Li, J. Li, T. Huang, X. Zhang, Y. Li, J. Wu, et al. Hydra-x: Native unified multimodal models with holistic visual tokenizers. *arXiv preprint arXiv:2606.13289*, 2026.

[100] G. Zhang, Z. Zhou, T. Hu, Z. Peng, Y. Zhang, Y. Chen, Y. Zhou, Q. Lu, and L. Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025.

[101] B. Zheng, N. Ma, S. Tong, and S. Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[102] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A Technical Appendices

A.1 Preliminaries

The diffusion model [27, 88, 99, 56] learns to generate data by reversing a noise corruption process. Following Rectified Flow [45], the forward process linearly interpolates between a clean data sample $\boldsymbol{z}_0 \sim \boldsymbol{p}_{\text{data}}$ and Gaussian noise $\boldsymbol{z}_1 \sim \mathcal{N}(0, I)$:

$$\boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, \tag{10}$$

where $t \in [0, 1]$ is a continuous timestep controlling the noise level. A neural network $\hat{\boldsymbol{v}}_\theta(\boldsymbol{z}_t, t)$ is trained to regress the velocity field $\boldsymbol{z}_1 - \boldsymbol{z}_0$ from the noisy sample $\boldsymbol{z}_t$ and timestep t, enabling iterative denoising from pure noise back to clean data at inference. The training objective is:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{z}_1, t}(\|(\boldsymbol{z}_1 - \boldsymbol{z}_0) - \hat{\boldsymbol{v}}_\theta(\boldsymbol{z}_t, t)\|^2). \tag{11}$$

A.2 Evaluation Metrics

To comprehensively evaluate the model performance, we utilize multiple metrics to validate the model on Verse-Bench and Sem100 in terms of video quality, audio quality, video-audio synchronization, and prompt following accuracy.

Video Quality. We leverage imaging quality (IQ) and dynamic degree (DD) from VBench [32] to validate the synthesized video quality. We use the pretrained aesthetic-predictor-v2-5 [15] to predict the aesthetic quality (AQ) [28]. We further evaluate the identity consistency (ID) by comparing the mean DINOv3 [64] representation between the first synthesized frame and the rest of the synthesized frames.

Audio Quality. We utilize the pretrained AudioBox [71] to evaluate the perceptual audio quality across production quality (PQ) and content usefulness (CU). To assess the accuracy of multi-speaker speech, we propose the multi-speaker word error rate (M-WER). Given the synthesized audio, we first use Qwen3-Omni [93] to segment the speech into multiple transcripts by speaker identity based on vocal characteristics, and then compute the word error rate [58] against the ground-truth transcript.

Video-Audio Synchronization. We utilize the Sync-C [11] and Sync-D [11] to measure the synchronization of lips with audio. We further use the pretrained Synchformer [33] to quantify the temporal misalignment between video and audio streams.

Prompt Following Accuracy. We utilize Gemini-3.1 [19] to evaluate the prompt following accuracy of joint video-audio generation models. The evaluation prompts are shown in Fig. 7.

A.3 Details of Dual Semantic Alignment Towers

In the dual semantic alignment towers, $\texttt{CMAttn}_{v\to a}(\cdot)$ and $\texttt{CMAttn}_{a\to v}(\cdot)$ require each query token from one modality to attend to temporally corresponding key tokens from the other modality. However, video and audio tokens have fundamentally different temporal densities. A video contains N keyframes sampled at FPS $= 6$, with each keyframe encoded into n_v spatial tokens, yielding $L_v = N \times n_v$ tokens in total. Audio is represented as M one-second chunks, with each chunk encoded into n_a tokens, yielding $L_a = M \times n_a$ tokens. Without positional encoding, the cross-modal attention is content-based and cannot distinguish temporally aligned tokens from misaligned ones.

To address this, we assign a continuous timestamp τ to each token based on its actual temporal position, and apply Rotary Position Embedding (RoPE) using these timestamps in the $\texttt{CMAttn}_{v\to a}(\cdot)$ and $\texttt{CMAttn}_{a\to v}(\cdot)$. This maps both modalities onto a shared continuous time axis, so that the attention scores naturally reflect temporal proximity.

Timestamp Assignment. Since the total video duration is M seconds (equal to M audio chunks), we assign each token a physical timestamp in seconds based on its position within the total duration. For video, the N keyframes are uniformly distributed across M seconds, so the i-th keyframe ($i = 0, \ldots, N-1$) corresponds to time $\tau_v^{(i,\cdot)} = i \cdot M/N$ seconds. Since the n_v spatial tokens within a single frame represent different spatial patches of the same temporal instant, they share the same timestamp. Thus, for the j-th spatial token of the i-th keyframe:

$$\tau_v^{(i,j)} = \frac{i \cdot M}{N}, \quad i = 0, \ldots, N-1, \;\; j = 0, \ldots, n_v - 1, \tag{12}$$

For audio, the m-th chunk ($m = 0, \ldots, M-1$) covers the time interval $[m, m+1)$ seconds. Unlike video spatial tokens that share the same temporal instant, the n_a tokens within a chunk are arranged in temporal order (each corresponding to a successive temporal position in the encoder output), so they receive evenly spaced timestamps within the chunk:

$$\tau_a^{(m,j)} = m + \frac{j}{n_a}, \quad m = 0, \ldots, M-1, \;\; j = 0, \ldots, n_a - 1, \tag{13}$$

RoPE Application. Given token embeddings $\boldsymbol{h} \in \mathbb{R}^D$ and its timestamp τ, we apply RoPE as:

$$\texttt{RoPE}(\boldsymbol{h}, \tau) = \boldsymbol{h} \odot \cos(\tau \cdot \boldsymbol{\omega}) + \texttt{rotate}(\boldsymbol{h}) \odot \sin(\tau \cdot \boldsymbol{\omega}), \tag{14}$$

where $\boldsymbol{\omega} \in \mathbb{R}^{D/2}$ is the frequency vector with $\omega_k = \frac{1}{\theta^{2k/D}}$ (θ is a base frequency hyperparameter). $\odot$ denotes element-wise multiplication and $\texttt{rotate}(\cdot)$ swaps adjacent pairs with sign flips following the standard RoPE.

The dual towers involve two symmetric cross-modal attention operations. Let $\hat{\boldsymbol{F}}_v = \texttt{CAttn}_v(\boldsymbol{Q}_v, \boldsymbol{H}_v)$ and $\hat{\boldsymbol{F}}_a = \texttt{CAttn}_a(\boldsymbol{Q}_a, \boldsymbol{H}_a)$ denote the intermediate features from the intra-modal cross-attention of each tower. Both CAttn operations execute first, and then the two towers exchange features via CMAttn. In the video tower $\texttt{CMAttn}_{a\to v}(\cdot)$, $\hat{\boldsymbol{F}}_v$ serves as queries and $\hat{\boldsymbol{F}}_a$ as keys/values. In the audio tower $\texttt{CMAttn}_{v\to a}(\cdot)$, the roles are reversed. Both directions share the same timestamp assignment, with RoPE applied to queries and keys using their respective timestamps:

$$\begin{aligned}\texttt{CMAttn}_{a\to v}(\hat{\boldsymbol{F}}_v, \hat{\boldsymbol{F}}_a) &= \texttt{Attn}\big(\texttt{RoPE}(\hat{\boldsymbol{F}}_v,\, \tau_v),\, \texttt{RoPE}(\hat{\boldsymbol{F}}_a,\, \tau_a),\, \hat{\boldsymbol{F}}_a\big),\\ \texttt{CMAttn}_{v\to a}(\hat{\boldsymbol{F}}_a, \hat{\boldsymbol{F}}_v) &= \texttt{Attn}\big(\texttt{RoPE}(\hat{\boldsymbol{F}}_a,\, \tau_a),\, \texttt{RoPE}(\hat{\boldsymbol{F}}_v,\, \tau_v),\, \hat{\boldsymbol{F}}_v\big).\end{aligned} \tag{15}$$

Notably, since audio has no spatial structure, spatial positional encoding is not applied in CMAttn. All n_v spatial tokens within the same video frame share the same timestamp, allowing them to attend uniformly to the same set of temporally corresponding audio tokens.

A.4 Semantic Blueprint Interpretability

We agree the planned tokens are continuous hidden states rather than human-readable event plans. However, "explicit" in our term refers to three structural properties that distinguish planned tokens from conventional learned features. First, each token has a verifiable semantic target with one-to-one correspondence to a concrete perceptual feature (SigLip2 for a specific keyframe, WavTokenizer for a specific audio chunk), supervised via position-wise MSE (Eq. 8). Second, the tokens are structurally decomposed by modality (separate video and audio sequences) and by time (per-keyframe, per-chunk), making the temporal and modal structure of the plan transparent by construction. Third, the planning is stage-separated from synthesis, occurring as an explicit reasoning step before denoising rather than being entangled within the denoising process. These properties make the blueprint "explicit", as each token's semantic content is inspectable by measuring its MSE against the ground-truth perceptual feature at that temporal position, as shown in Table 9. The training mechanism further guarantees that the MLLM must perform genuine semantic reasoning. Consider a prompt "a man swings a bat, the ball flies into the air, and a crowd cheers." To minimize the MSE at each temporal position, the MLLM must internally decompose this into the correct temporal sequence (keyframe 1 encodes the swing posture, keyframe 2 encodes ball contact, keyframe 3 encodes the ball in flight) with corresponding audio (chunk 1 captures the swing whoosh, chunk 2 captures the impact crack, chunk 3 captures crowd cheering). If the MLLM fails to reason about temporal ordering or causal relationships, the predicted features at each position would mismatch the ground truth. The training objective therefore directly enforces semantic reasoning as a hard supervised requirement.

We acknowledge that these tokens also function as learned conditioning features for the DiT, and this is by design rather than in contradiction. The two roles are complementary. The Stage-2 in Table 8 confirms that ground-truth SigLip2/WavTokenizer features achieve P-Acc=0.87, while our model achieves 0.82, with the gap stemming from imperfect prediction accuracy. This proves the bottleneck is semantic content quality, not the conditioning pathway, further justifying the "semantic blueprint" characterization.

A.5 Dataset Details

Regarding the training dataset, our training dataset (1.5 million video-audio clips, FPS=24) is aggregated from OpenHuman-Vid [42], AudioCaps [36], WavCaps [51], and videos collected from the internet. We apply a multi-stage filtering pipeline to ensure data quality. First, we use aesthetic-predictor-v2-5 [15] to predict the visual aesthetic quality of each video and retain only clips with a score above 0.4. We then filter out static or near-static videos by requiring a Dynamic Degree [32] above 0.2. For audio quality, we use AudioBox [71] to evaluate audio aesthetics and keep only clips with PQ above 6.0. Finally, for videos containing speech, we apply SyncNet [11] to assess lip-sync accuracy and retain only clips with a confidence score above 0.9. The resulting dataset covers a diverse range of audiovisual scenarios, including single-speaker and multi-speaker speech or singing, natural environmental sounds (e.g., wind, rain, animal calls), sci-fi and virtual environment sounds, and human-object or human-environment interaction scenes (e.g., cooking and instrument playing). Furthermore, we utilize Qwen3-VL-235B-A22B-Instruct [2] and Qwen3-Omni [93] to caption the training video dataset to gain the video prompts and audio prompts, respectively.

In terms of the testing dataset, we select 100 unseen videos (10 seconds long, FPS=24) from the internet to construct the testing dataset Sem100. Some ground truth examples are shown in Fig. 8.

Table 5: Quantitative comparison results between Baton and commercial models.

Model	AQ↑	DD↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑
Veo3.1 [18]	0.58	0.53	8.27	7.96	0.14	0.35	0.75
Wan2.7 [1]	0.63	0.62	8.18	7.75	0.19	0.38	0.85
Kling3.0 [40]	0.64	0.56	8.12	7.82	0.13	0.45	0.72
Seedance2.0 [7]	**0.68**	**0.68**	**8.54**	**8.07**	**0.11**	**0.28**	**0.92**
Ours	0.54	0.48	7.53	7.16	0.14	0.68	0.82

The sources of videos come from numerous social media platforms, including YouTube and BiliBili. These videos feature individuals across diverse ethnicities, genders, and age groups, portrayed in full-body, half-body, and close-up shots against varied indoor and outdoor settings. Sem100 covers a broad spectrum of audiovisual scenarios, including single-speaker and multi-speaker dialogues, singing performances, natural environmental sounds, human-object interactions (e.g., cooking, sports, musical instruments), and human-environment interactions in both realistic and stylized scenes. In contrast to existing open-source testing datasets (Verse-Bench), the text prompts of Sem100 are deliberately designed to demand complex semantic reasoning, involving multi-stage sequential actions, compositional instructions, and intricate causal relationships across both visual and auditory modalities. The full text prompts are saved in the JSON file in the supplementary material (zip). The licenses of Verse-Bench and Sem100 are both Apache-2.0.

A.6 Training Details

We use the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. The model is trained at bf16 precision, equipped with DeepSpeed-Stage-3 for distributed data parallel training. The training process contains three stages, and the learning rates in all stages are $1e-5$. The learnable queries $\boldsymbol{Q}_v \in \mathbb{R}^{L_v \times D}$ and $\boldsymbol{Q}_a \in \mathbb{R}^{L_a \times D}$ in the dual semantic alignment towers are initialized from a normal distribution ($\sigma = 0.02$) and have the same sequence length as the planned video and audio tokens (L_v and L_a), respectively. The training video dataset in the 3 training stages is the same. Fig. 6 shows the MLLM input template during training. In particular, in the first training stage (VA-Planner Pretraining), the MLLM is initialized by Qwen3-8B [95], and the whole MLLM and $\texttt{SMLP}_{v/a}(\cdot)$ are trainable. The frozen video semantic encoder is initialized by SigLip2 [72] (siglip2-so400m-patch14-384), and the frozen audio semantic encoder is initialized by WavTokenizer [34] (WavTokenizer-large-unify-40token). We train the MLLM for 10 epochs on 160 NVIDIA H800 GPUs, with a batch size of 1 per GPU. In the second training stage (DiT Semantic Adaptation), the DiT is initialized by Ovi [48]. DiT and $\texttt{LMLP}_{v/a}(\cdot)$ remain trainable. We train the DiT for 10 epochs in 640p videos (FPS=24) on 160 NVIDIA H800 GPUs, with a batch size of 1 per GPU. In the third training stage, only DiT remains trainable, and MLLM is frozen. We still train the DiT for 10 epochs in 640p videos (FPS=24) on 160 NVIDIA H800 GPUs, with a batch size of 1 per GPU.

A.7 More Comparison Results

Fig. 9, Fig. 10, and Fig. 11 show additional comparison results between our Baton and the state-of-the-art open-source models. While previous models [48, 100, 21, 70] often struggle with user prompts involving complex sequences of actions, multi-person dialogue, and intricate interactions between characters and their environment, they tend to misunderstand or omit key narrative details. As a result, the generated audio and video may not align with the prompt or may even be unsynchronized with each other. In some cases, discrepancies in how the audio and video branches interpret coarse global text embeddings can lead to conflicts during joint generation, causing visual artifacts such as body distortion or audio issues like noise and unnatural effects. In contrast, Baton accurately follows user prompts and produces stable, high-quality, and well-synchronized video-audio outputs.

A.8 Comparison with Commercial Models

Table 5 and Fig. 12 depict the comparison results between our Baton and closed-source commercial models. Although Baton lags behind commercial models in visual quality and audio aesthetics, it achieves comparable performance in prompt following. For instance, in Fig. 12, the right-side case shows that Veo 3.1 generates an incorrect number of people, Kling 3.0 and Wan 2.7 fail to match the specified positions of the pouring character, and Seedance 2.0 produces an incorrect order of

Table 6: Ablation study on the injection of planned tokens. Concat with Text Embeddings refers to concatenating the planned tokens with the global text embeddings. Parallel TCAttn+V/ACAttn performs TCross-Attn and V/ACross-Attn in parallel, rather than the original cascaded design.

Model	AQ↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑
Concat with Text Embeddings	0.50	7.45	7.01	0.17	0.77	0.76
Parallel TCAttn+V/ACAttn	0.52	7.48	7.14	0.14	0.71	0.79
Ours	**0.54**	**7.53**	**7.16**	**0.14**	**0.68**	**0.82**

Table 7: Ablation study on keyframe FPS in the planning stage.

Model	AQ↑	DD↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑	GPU Mem↓
FPS=1	0.44	0.36	7.18	6.81	0.46	0.78	0.70	59.6G
FPS=4	0.47	0.41	7.35	6.92	0.33	0.72	0.78	62.7G
FPS=6 (Ours)	0.54	0.48	7.53	7.16	0.14	0.68	0.82	68.4G
FPS=8	**0.57**	**0.49**	**7.81**	**7.23**	**0.12**	**0.59**	**0.85**	**74.2G**
FPS=12	-	-	-	-	-	-	-	OOM

Table 8: Ablation on Orders of T_v^{tag} and T_a^{tag}.

Model	AQ↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑
Interleave <|img_pad|> and <|aud_pad|> [93]	0.47	7.24	6.83	0.31	0.75	0.75
$[T_a^{tag}; T_v^{tag}]$	0.50	7.31	6.97	0.25	0.73	0.79
$[T_v^{tag}; T_a^{tag}]$ (Ours)	**0.54**	**7.53**	**7.16**	**0.14**	**0.68**	**0.82**

Table 9: Ablation study on training stages. Stage-2 directly utilizes the groundtruth SigLip2 embeddings and WavTokenizer embeddings as the planned tokens, and injects them into DiT without MLLM prediction. Skip Stage-2 refers to omitting Stage-2 training, and the training process includes only Stage-1 and Stage-3.

Model	AQ↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑
Stage-2	**0.60**	**7.92**	**7.70**	**0.12**	**0.45**	**0.87**
Skip Stage-2	0.49	7.26	6.73	0.40	0.82	0.75
Ours	0.54	7.53	7.16	0.14	0.68	0.82

character appearances. By contrast, Baton faithfully generates video-audio content that aligns with the text prompt, demonstrating its superiority in handling complex prompts that demand semantic reasoning.

A.9 More Ablation Study

Ablation on Planned Token Injection. We additionally conduct an ablation study on the injection of planned tokens predicted by our VA-Planner. The results are depicted in Table 6. We can see that our injection achieves the best performance. Concatenating planned tokens with text embeddings forces both signals to share the same attention layer, where the localized spatio-temporal semantics of the planned tokens are diluted by the coarse global text signal, degrading DeSync and P-Acc. In parallel injection, both signals independently modify the raw latents without hierarchical interaction, losing the chance for semantic complementarity. Our cascaded design first lets text cross-attention establish a coarse semantic prior, then V/ACross-Attn refines it with fine-grained, position-specific planning cues, yielding the best performance.

Ablation on Keyframe FPS in Planning Stage. We ablate the keyframe FPS in the planning stage, as shown in Table 7. We observe that as the keyframe FPS increases, more keyframe blueprints are provided to the DiT, leading to improved performance. However, GPU memory consumption also rises significantly. At FPS = 12, single-GPU inference exceeds 80GB and results in Out-Of-Memory (OOM) errors. Notably, the performance gain from FPS = 8 over FPS = 6 is smaller than that from FPS = 4 to FPS = 6. We attribute this to semantic saturation. At FPS = 6, the keyframe-level semantic cues provided to the DiT are already sufficient, and further increasing the number of keyframes introduces redundant information rather than additional useful semantics. To balance computational cost and model performance, we therefore choose FPS = 6 as the final configuration.

Table 10: Ablation study on planned token prediction accuracy.

Model	AQ↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑	MSE Loss↓
w/o Learnable Query	0.52	7.25	6.88	0.27	0.74	0.79	0.37
w/o Tower	0.48	6.92	6.53	0.45	0.78	0.69	0.39
w/o RoPE in Tower	0.42	6.57	6.24	0.68	1.05	0.44	0.46
Baton-Qwen3-VL-8B-Instruct	0.44	6.90	6.51	0.45	0.84	0.72	0.45
Baton-Qwen3-Omni-30B-A3B	0.41	6.76	6.28	0.65	1.02	0.65	0.50
Baton-Qwen3-4B	0.46	7.32	6.90	0.31	0.82	0.68	0.41
Baton-Qwen3-32B	**0.58**	**7.86**	**7.29**	**0.13**	**0.54**	**0.85**	**0.27**
Baton-Qwen3-30B-A3B-Instruct	0.55	7.51	7.12	0.16	0.61	0.82	0.35
Ours (Baton-Qwen3-8B)	0.54	7.53	7.16	0.14	0.68	0.82	0.31

Ablation on Orders of T_v^{tag} and T_a^{tag}. We ablate the position orders of T_v^{tag} and T_a^{tag} in the MLLM input, as shown in Table 8. Interleave <|img_pad|> and <|aud_pad|> [93] follows the Qwen3-Omni [93], which interleaves <|img_pad|> and <|aud_pad|> at the chunk level so that temporally co-occurring video and audio tokens are adjacent. Despite providing explicit temporal proximity, interleaving fragments each modality's token sequence, preventing the MLLM from building coherent intra-modal representations before cross-modal reasoning, and yields the worst results. $[T_a^{tag}; T_v^{tag}]$ places audio before video, which improves over interleaving but still underperforms Ours. In the causal attention of MLLM, audio tokens can only condition on the text prompt without any visual context, so the audio plan is generated blind to the visual scene. The subsequent video tokens do attend to the audio plan, but this inverts the natural dependency. In most scenarios, sounds accompany and react to visual events rather than the reverse. Our $[T_v^{tag}; T_a^{tag}]$ ordering exploits this asymmetry by letting the MLLM first build a complete video plan from the text prompt, then conditioning audio planning on both the prompt and the full video plan via causal attention. This produces audio tokens that are inherently synchronized with the visual blueprint.

Ablation on Training Stages. We conduct an ablation study on training stages, as shown in Table 9. The results indicate that Stage 3 acts as a critical adaptation phase, aligning training conditions with inference-time planner outputs. By replacing ideal encoder features with realistic predictions, it reduces exposure bias and significantly improves robustness during generation.

Ablation on Planned Token Prediction Accuracy. We conduct an ablation study on planned token prediction accuracy, as shown in Table 10. In particular, for the MSE Loss column, we take each Stage-1-trained MLLM variant and run it on the training set. We extract ground-truth semantic embeddings from frozen SigLip2 and WavTokenizer encoders and compute the MSE against the MLLM-predicted planned tokens, thereby measuring how accurately the VA-Planner learns to approximate the target perceptual features after Stage-1 pretraining. All other metrics are evaluated on Sem100 using the full three-stage model. The results reveal a strong positive correlation between Stage-1 prediction accuracy (MSE) and final generation quality. Qwen3-32B achieves the lowest MSE and the best downstream metrics, confirming that stronger language model reasoning capacity produces more accurate planned tokens, which in turn provides higher-quality blueprints for the DiT. Conversely, *w/o* RoPE in Tower and Qwen3-Omni exhibit the high MSE, and correspondingly suffer the most severe degradation in P-Acc and DeSync, indicating that temporally misaligned or distribution-biased planned tokens mislead rather than guide the denoising process. Notably, the correlation is not strictly monotonic. Ours achieves comparable metrics to Baton-Qwen3-30B-A3B despite similar prediction accuracy. Since MSE only measures pointwise token-level regression error, it cannot capture the inter-token temporal coherence across the planned token sequence. As discussed in the backbone ablation, MoE's sparse routing activates different experts per token, fragmenting features and weakening holistic cross-token reasoning. Consequently, even when the average per-token MSE is similar, the planned tokens from MoE may lack the globally consistent temporal structure that a dense model naturally maintains through unified parameter sharing, and this structural coherence is critical for guiding long-horizon denoising. Furthermore, since the planned token prediction is not perfect, Stage-3 joint fine-tuning can effectively compensate for planner imperfections by adapting the DiT to realistic prediction noise. However, it cannot rescue severely inaccurate plans (e.g., *w/o* RoPE in Tower), highlighting that a minimum level of Stage-1 prediction quality is necessary for the three-stage curriculum to succeed.

Table 11: Inference speed and GPU memory consumption comparison results.

Model	AQ↑	PQ↑	CU↑	M-WER↓	DeSync↓	P-Acc↑	GPU Mem↓	Speed↓
Ovi [48]	0.42	6.74	6.23	0.66	1.06	0.46	53.6G	184s
UniAVGen [100]	0.31	6.68	5.83	0.72	1.14	0.44	46.8G	306s
LTX-2 [21]	0.48	7.14	6.91	0.58	0.97	0.62	48.5G	196s
MOVA [70]	0.44	6.95	6.94	0.61	1.05	0.47	76.7G	1517s
Ours	**0.54**	**7.53**	**7.16**	**0.14**	**0.68**	**0.82**	68.4G	346s

Table 12: User preference of Baton compared to other competitors. A higher score indicates users prefer more to our model.

Model	V-A	A-A	I-C	B-C	V-A-S
Ovi [48]	94.8%	95.3%	94.1%	94.4%	93.5%
JavisGPT [46]	95.2%	96.4%	95.0%	94.9%	96.6%
UniAVGen [100]	96.7%	97.2%	98.3%	95.5%	96.2%
LTX-2 [21]	90.4%	88.6%	91.8%	90.2%	89.4%
MOVA [70]	93.5%	94.2%	91.6%	89.4%	91.0%

A.10 Speed and GPU Resource Comparison

We compare the inference speed and GPU memory consumption between our Baton and the previous joint video-audio generation models, as shown in Table 11. Notably, we utilize the full version of LTX-2 [21] instead of the distilled version for fair performance comparison. Compared to our DiT backbone Ovi [48], Baton introduces approximately 28% additional GPU memory from the VA-Planner's autoregressive planning and extra cross-attention layers, while achieving over 36% improvement in DeSync and 78% in P-Acc, demonstrating that the planning overhead is a highly favorable trade-off. Compared to MOVA [70], Baton uses 11% less GPU memory and is $5\times$ faster while delivering substantially better quality across all metrics, indicating that explicit semantic planning is a more efficient path to cross-modal coordination than simply scaling model and data.

A.11 User Study Details

Table 12 shows the user study results. Fig. 20 illustrates a screenshot of our user study. The 30 test videos were randomly sampled from our Sem100 dataset, which consists of 100 unseen videos (10 seconds long) collected from diverse social media platforms (YouTube, BiliBili). These videos span various ethnicities, genders, and indoor/outdoor settings.

A total of 200 individuals took part in this evaluation. Eligibility was restricted to adults aged 18 and older who possessed normal or corrected-to-normal visual acuity and reported no auditory deficits. Participants were recruited through internal university mailing lists and campus announcements at our institution. The participant pool drew predominantly from the university's students (approximately 75%), with the remainder comprising faculty members (25%). Disciplinary backgrounds varied widely, including computer science, engineering, and the arts. The sample maintains an approximately even gender balance. The evaluation followed a two-alternative forced choice (2AFC) paradigm. On every trial, text prompts are first displayed to the participant, after which two videos appear in a randomized sequence: one produced by Baton and the other by a competitor. Participants indicated which of the two videos they judged superior along five criteria: video alignment with text prompts, audio alignment with text prompts, identity consistency, background consistency, and video-audio synchronization. Every participant assessed the full set of 30 cases, with the left–right ordering of the two clips randomized independently per trial to mitigate positional bias. All responses are gathered through an online survey interface. Preference rates are then computed for every method pairing by aggregating judgments across the entire participant pool and all cases. In terms of IRB approval, we confirm that we consulted our university's institutional ethics board before conducting the study. The study was granted an exemption from full IRB review under the institution's minimal-risk research provisions, based on three conditions. First, the study involved only the perceptual comparison of pre-generated video outputs and did not involve any intervention, deception, or interaction beyond viewing videos and selecting preferences. Second, no personally identifiable information (such as names, email addresses, or demographic details beyond aggregate statistics) was collected or stored. Third, participation was entirely voluntary, and participants were informed of the study's purpose and

their right to withdraw at any time before beginning. Each participant generally takes 10 minutes to accomplish our user study, and has received 5 dollars for compensation.

A.12 Multi-Speakers Results

Fig.13 shows some multi-speaker cases. The results demonstrate that Baton is capable of handling complex scenarios involving multi-person interaction or communication.

A.13 Cartoon Video Results

To validate the diversity of Baton, we experiment on a cartoon case, as shown in Fig. 14. The results indicate that Baton can also synthesize cartoon or virtual videos, highlighting its superiority in diversity.

A.14 Complex Scene Results

Fig. 15, Fig. 16, Fig. 17, Fig. 18, and Fig. 19 show the complex scene results. The test cases encompass complex scenarios requiring semantic reasoning, including sequences of multiple narrative events, continuous interactions between characters and their environment or objects, logically coherent camera movements, large-scale interactions (motion) among multiple characters, evolving natural scenes, ordered multi-event actions involving specific individuals, and environment-interaction sounds that adhere to physical laws. For example, the first, second, and fourth rows in Fig. 15 demonstrate narrative-driven character-environment interactions with physically plausible ambient sounds that respect the acoustic properties of the scene. The third row in Fig. 15 and the fourth row in Fig. 16 showcase ordered multi-event sequences involving specific individuals, where each action must be completed before the next begins in the correct temporal order. The fifth row in Fig. 16 and the first row and the third row in Fig. 17 present complex narrative-driven human-object interactions that require reasoning over causal dependencies between actions and their outcomes. The first row in Fig. 16 and the fourth and fifth rows in Fig. 17 feature logically coherent camera movements that follow the narrative progression of the scene. Fig. 19 contains cases with intense multi-character interaction dynamics, instructional narrative descriptions, and interpersonal interactions between characters. We can observe that our Baton can perform a wide range of joint video-audio generation while accurately following the complex user prompt description and preserving the protagonist's appearance, background, identity, and synchronization between video and audio.

A.15 Human Subjects Data Concern

Our training and testing datasets both involve human subjects, containing identifiable information. Since the collected videos are primarily from BiliBili and YouTube, we contacted the video uploaders via private messages to inquire whether they agreed to allow their videos to be used for non-commercial academic research, and we obtained their consent. All videos in our training and testing datasets have been approved by their corresponding video uploaders.

A.16 Ethics Concerns

Our Baton can synthesize joint video-audio content based on the given text prompts, which can be implemented in digital human creation and film creation. However, Baton carries risks of misuse, including the creation of deepfake videos for identity impersonation, non-consensual manipulation of a person's likeness, and the spread of misinformation through fabricated speech videos on social media. To mitigate this, it is essential to integrate visible and invisible watermarking into all generated content to ensure traceability during public deployment, incorporating automated sensitive content detection and moderation pipelines (e.g., deepfake detectors) before any output is released, restricting access through API-level authentication and usage logging, and establishing clear terms of service that prohibit generating content of real individuals without their explicit consent.

A.17 Limitations and Future Work

Fig. 21 shows one failure case of our Baton. When generating scenes involving multiple characters where each individual occupies a relatively small spatial region in the frame, the synthesized faces

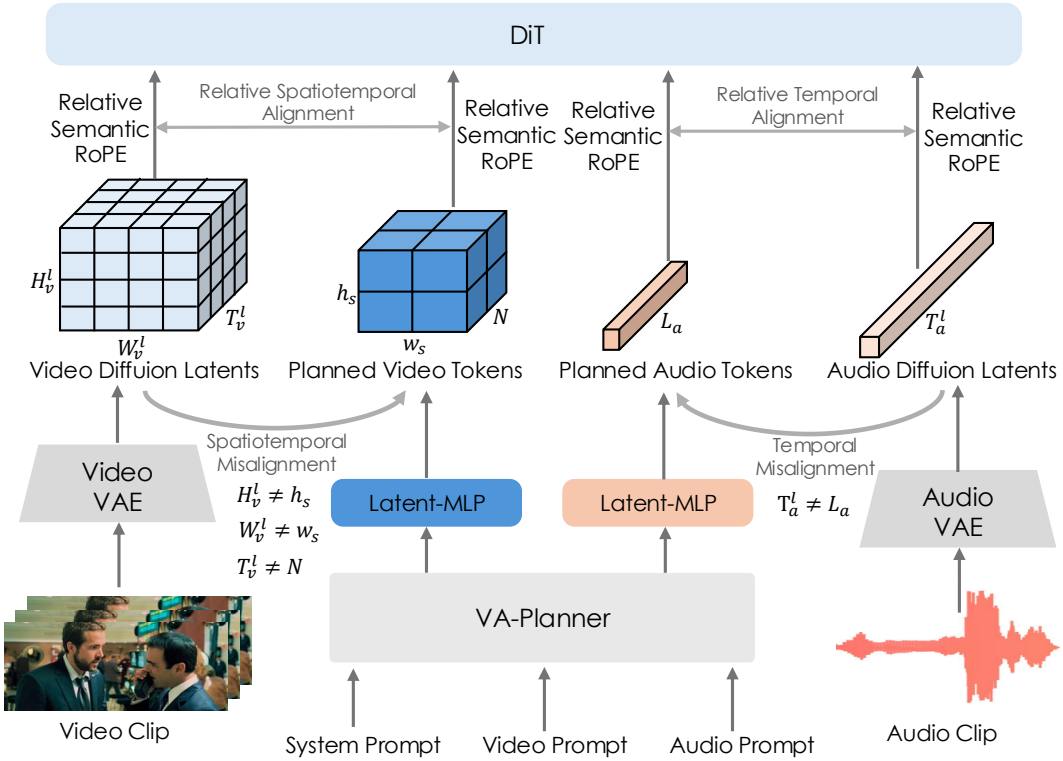


Figure 4: The motivation of Relative Semantic RoPE).

tend to appear blurry and lack fine-grained details. This is because the planned tokens are predicted at a sparse keyframe resolution with limited spatial tokens per frame, and when characters are small, only a few spatial tokens cover each face region, providing insufficient semantic guidance for the DiT to reconstruct detailed facial features. Moreover, the underlying 3D VAE compresses spatial dimensions aggressively, further reducing the effective resolution available for small faces. One potential solution is to incorporate a dedicated face refinement module or a super-resolution stage that operates on detected face regions, or to adopt an adaptive spatial token allocation strategy that assigns denser planned tokens to regions containing small faces. This part is left as future work.



Figure 5: Ablation study on VA-Planner (a), RS-RoPE (b), and different backbones (c).

Baton MLLM Input Template

```
conversation = [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": f"Generate {N} key frames and {M} seconds corresponding audio sequences based on the video
                  instruction: {Video Prompt} and audio instruction: {Audio Prompt}."
    },
    {
      "role": "assistant",
      "content": "<|v_start|>" + "<|img_pad|>" * L_v + "<|v_end|>" +
                 "<|a_start|>" + "<|aud_pad|>" * L_a + "<|a_end|>"
    }
]
```

Figure 6: The MLLM input template (system prompt, video prompt, and audio prompt).

Gemini Evaluation System Prompt

You are an expert audiovisual content evaluator. Your task is to analyze a generated video with its accompanying audio based on a video prompt and an audio prompt.

Please provide an overall Prompt-Following Accuracy (P-Acc) score between 0.0 and 1.0, along with a brief analysis. You must comprehensively consider all of the following aspects when determining the score.
Video Prompt: "{Video Prompt}"
Audio Prompt: "{Audio Prompt}"

Aspects to Consider
1. ** Visual Prompt Alignment **
   Whether the video accurately depicts all actions, interactions, and motion sequences described in the video prompt, performed by the correct subjects in the correct temporal order. Whether the camera movement and framing match any cinematographic instructions (e.g., fixed shot, panning, close-up). Whether the correct number of characters appear at the right moments and in the right order. Whether multi-stage actions are fully completed without omission or reordering. Whether the scene setting, subjects, objects, and their visual attributes (appearance, color, clothing, spatial layout) match the description, and whether subjects remain visually consistent throughout without identity drift or body distortion.

2. ** Audio Prompt Alignment **
   Whether the audio contains the correct sound events, ambient sounds, music, or speech described in the audio prompt with matching types, timbres, and characteristics. Whether speech or dialogue has the correct spoken content, speaker identity, language, and tone. Whether the generated sounds obey physical plausibility (e.g., footsteps on wood sound different from footsteps on gravel, a heavy object produces a louder impact than a light one, echo in a large room).

3. ** Human-Object and Human-Environment Interaction **
   Whether characters interact with the correct objects in the correct manner as described (e.g., picking up the right item, placing it at the right location). Whether character-environment interactions are physically plausible (e.g., feet touching the ground while walking, hands making contact with surfaces during interactions).

4. ** Joint Video-Audio Alignment **
   Whether audio events are temporally synchronized with their corresponding visual events (e.g., a clap sound aligns with hand contact, footsteps align with foot movement, speech aligns with lip movement). Whether the audio and video semantically correspond as a coherent audiovisual experience (e.g., piano sound when a person plays piano, not when they are walking). Considering both prompts together as a unified instruction, whether the generated video-audio content holistically fulfills the combined intent without gaps or contradictions between modalities.

Based on your comprehensive evaluation of the above aspects, provide a overall P-Acc score. Please respond in this exact JSON format:
```
{
  {
    "analysis": "A brief paragraph summarizing the strengths and weaknesses across all aspects.",
    "P-Acc score": X.X
  }
}
```

Figure 7: The system prompt used for Gemini-based prompt following accuracy evaluation.



Figure 8: Examples from Sem100.



Video Prompt: Against a red brick wall, a male soldier with short dark hair in a greenish-brown military jacket leans toward a female companion on his right, who has styled dark hair under a brown beret and wears a brown coat with a thick fur collar. He moves in to kiss her, but she gently pushes him back with her hand on his shoulder and speaks earnestly while looking down at him. They pause, pressing their noses together, then share a tender kiss. He pulls away and begins to move off; she watches him go, her expression one of quiet contentment. Cinematic realism with low saturation and a cool color temperature overall. Intimate. The scene takes place outdoors during daytime, likely near a building made of red bricks. The environment is a historical period in World War I. Other figures can be seen walking around in the out-of-focus background, creating a sense of activity but also emphasizing isolation. The lighting is natural daylight.

Audio Prompt: A soft, melancholic instrumental music piece plays, featuring gentle piano melodies and sustained strings, creating a somber and emotional atmosphere. A young woman [Speaker A], her voice trembling with emotion and urgency, speaks in a breathy, pleading tone: \"Come back to me." She continues, her voice breaking slightly: \"Stay alive and come back to me." The music swells subtly, underscoring the emotional weight of her words before fading out.

Video Prompt: On a sunlit outdoor asphalt basketball court, bordered by dense green trees under a clear blue sky, a young man with short brown hair wearing dark sunglasses, a grey baseball cap, a black t-shirt and black Nike athletic shorts. He picks up a red basketball. He stands, turns away from the camera, and walks along the baseline, dribbling the ball between his legs. As he nears the free-throw line he takes a jump shot; the ball arcs over the rim and drops through the net. Medium-to-tracking shot, eye-level, with leading room guiding the viewer toward the basket. Natural high-key golden-hour daylight casts soft shadows across the grey court. Observational tone capturing the action in real time.

Audio Prompt: Set in an outdoor environment with birdsong and faint rustling sounds, a young man [Speaker A] speaks in a calm, encouraging tone: \"Easy peasy, baby.\" The sound of a ball being dribbled on a hard surface is heard, followed by a sharp impact as it hits a backboard or wall. The dribbling resumes, accompanied by the soft thud of the ball bouncing on the ground.

Figure 9: More comparison results (1/3). Please refer to the demo video for audio.



Video Prompt: From a low angle inside a dark trench, a young white soldier in his late teens wearing a Brodie helmet, a thick brownish-green jacket, webbed ammunition pouches, fingerless knit gloves, and a large backpack grips a Lewis Gun with both hands and cautiously pushes up the steep earthen steps. His dirt-smudged face shows intense concentration as he aims steadily toward a damaged stone building above under an overcast sky. Suddenly, a single gunshot cracks from the building above; the round narrowly misses him and strikes the top step, sending a puff of dirt into the air. Startled, he flinches and stumbles backward down the steps. He collapses heavily against the rough concrete trench wall, mouth agape in a silent scream of shock. His expression shifts to dazed disbelief as he looks upward, breathing heavily with wide eyes. Diffused overcast light casts soft shadows in a desaturated, cool-toned palette of muted greens, greys, and browns. Rubble-strewn ground and a blown-out brick building in the background complete a grim, battle-scarred wartime setting of gritty realism.

Audio Prompt: An outdoor ambient atmosphere with faint wind. Steady footsteps climb rough stone steps. Suddenly, a single sharp rifle shot cracks and echoes across the open air. After a brief silence, a soldier [Speaker A] breathes heavily in rapid, shaky gasps that gradually slow into exhausted, labored exhales.

Video Prompt: Indoors against a light gray wall, a large projection screen behind displays a bold red logo with a white bone icon above the words "BARK DOG ADVERTISING." On the left stands an older Black man with short dark hair in a charcoal grey pinstripe suit jacket, deep red collared shirt, red pocket square. To his right stands a middle-aged white woman with shoulder-length wavy brown hair and bangs, wearing a leopard-print blouse with billowy sleeves, a high-waisted black leather skirt, layered gold necklaces. The woman speaks first, raising her right index finger to emphasize a point. The man then responds, lifting both palms up in an explanatory gesture. The woman speaks again. The man replies, bringing his hands together in front of him as he talks. The woman finishes the exchange, lowering her hands with a concluding nod.

Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a mature woman [Speaker A] speaks in a steady, confident tone: \"I both worked in creative at Ogilvy and were very successful.\" A mature man [Speaker B] responds in a similarly confident, slightly boastful tone: \"Hugely successful. Shaq and the General Insurance.\" The woman [Speaker A] interjects proudly: \"Ours.\" The man [Speaker B] continues: \"Shaq and Icy Hot.\" The woman [Speaker A] replies with equal pride: \"Ours.\"

Figure 10: More comparison results (2/3). Please refer to the demo video for audio.



Video Prompt: On a sandy ocean floor with purple and green rock formations, stylized coral, and glowing ethereal patterns on a dark blue cliff face behind, Patrick — a pink starfish in a green shirt and shorts — sits on a wooden log looking sad. SpongeBob — a yellow sea sponge in his white shirt, brown pants, and small hat— stands nearby to the right, looking concerned. Patrick suddenly bursts into loud crying, two massive streams of tears pouring from his eyes like waterfalls. The force of his own crying pushes him backward, rolling him off the log. He quickly gets back up, drapes himself over the log, and continues sobbing. Then he suddenly lifts his head — his two streams of tears intensify into powerful blue laser-like beams that blast directly at SpongeBob, launching him off-screen. SpongeBob's small hat stays behind, spinning in mid-air before gently dropping to the sandy ground. Medium shot, eye-level, rule-of-thirds composition with Patrick on the left third and SpongeBob on the right. High-key, even lighting enhances the vibrant saturated palette of pink, yellow, green, and deep blues. 2.5D animation style with flat character designs rendered against depth-layered backgrounds, exaggerated proportions, fantastical physics, and a comedic, absurd tone.

Audio Prompt: In a chaotic and noisy environment, a young boy [Speaker A] begins to speak in a high-pitched, excited voice: \"Hi, my name is...\" when she is abruptly cut off by a loud, high-pitched scream. A man [Speaker B] lets out a long, drawn-out cry of pain or surprise. The boy [Speaker A] tries again, saying, \"Hi, my name is...\" but is immediately interrupted by another piercing scream. The audio is filled with overlapping screams and chaotic background noise throughout.

Video Prompt: A young Caucasian man stands at an outdoor shooting range, holding a scoped AR-15 rifle, he fires several shots at a nearby pine tree, then reloads.

Audio Prompt: In a quiet, open outdoor environment, a sharp gunshot rings out, followed by a male voice [Speaker A] saying \"Ah\" in a neutral tone. Immediately after, another gunshot is fired. After a brief pause, a mechanical click is heard, as if a weapon is being reloaded.

Figure 11: More comparison results (3/3). Please refer to the demo video for audio.



Video Prompt: Inside a modern, multi-level space agency headquarters with concrete pillars and stone floors, a military officer in an air force dress uniform addresses a gathered team of scientists in white lab coats, gesturing emphatically. He finishes, turns, and strides down a corridor continuing to speak with animated gestures. He enters an office flanked by two security guards, revealing a vast workspace of desks and data monitors, with a festive holiday display of lit miniature trees glowing inside against floor-to-ceiling windows framing snow-covered mountains. Medium shots on groups transition to tracking shots following the officer; eye-level, soft diffused ambient light with warm holiday string lights, cinematic realism style.

Audio Prompt: In a festive indoor setting with a low-level room tone, a mature man [Speaker A] speaks in a cheerful, enthusiastic tone: \"So we thought, might be a good idea to gather everyone together, and wish you a merry Christmas!" As he finishes, a burst of upbeat, orchestral holiday music swells, featuring bright brass and chimes. A chorus of voices [Speaker B] respond in unison with a joyful, celebratory tone: \"Merry Christmas!" The music continues as [Speaker A] speaks in a playful, commanding tone: \"Let us march!" The music swells again, becoming more prominent and triumphant as the scene concludes.

Video Prompt: At a nighttime high-rise terrace, nine people sit around a long dining table laden with food and wine glasses. At the center of the table sits a bearded man in a brown turtleneck, who stares wide-eyed with his mouth agape in astonishment. To his left, four guests are seated in order: a woman in a green dress, one men in a dark floral shirt, one man in a black shirt, and a woman in a white camisole. To his right, four more guests are arranged: a curly-haired man in a black long-sleeved shirt, a man in a casual jacket, a woman in a white camisole, and a woman in a gold camisole. The camera begins in a close-up on the bearded man at center, then the curly-haired man to his right holds a dark wine bottle and magically pours wine into a glass, astonishing the bearded man. As the camera slowly pulls back, the remaining guests gradually appear on screen from the center, one man raises his glass of red wine to propose a toast, and the others follow, lifting their glasses one by one. All nine guests break into smiles, lean in, and clink their glasses together over the center of the table in a joyous collective toast.

Audio Prompt: In a lively indoor setting with the ambient murmur of a social gathering, a man [Speaker A] with a calm, mid-range voice suggests, \"Un brindis. Un brindis y ya." A second man [Speaker B], with a slightly higher-pitched and more insistent tone, replies, \"No." The third man [Speaker C] then encourages, \"Venga, va. Un brindis por el futuro." A woman [Speaker D] with a clear, mid-range voice echoes, \"Por el futuro." The fourth man [Speaker E] affirms, \"Eso es. Por el futuro." The group [Speaker F] responds with a collective, low-pitched \"¡Oh!" as the sound of glasses clinking together is heard.

Figure 12: Comparison results between Baton and commercial models. Please refer to the demo video for audio.



Video Prompt: On an outdoor residential patio during an overcast day, with green grass, trees, a house facade in the background, and clusters of pink and white balloons creating a festive atmosphere, a middle-aged man with dark hair in a brown double-breasted suit, white shirt, and patterned tie stands facing a boy wearing a light brown suit jacket over a white shirt with a brown-and-black striped tie. Between them sits a table piled with wrapped gifts and chairs draped in leopard-print cloth. The middle-aged man speaks seriously, looking into the boy's eyes, then gestures emphatically toward the boy. The boy stands still and attentive. Naturalistic drama with a tense tone.

Audio Prompt: In a outdoor, a mature man [Speaker A] speaks in a inquisitive tone: \"What options" A younger man [Speaker B] responds in a neutral tone: \"Other offers." The first man [Speaker A] replies with a tone of mild surprise: \"You've had other offers" The second man [Speaker B] states in a neutral tone: \"Travis from Channel 87 wants to hire me full-time." The first man [Speaker A] asks in a slightly incredulous tone: \"To do what"



Video Prompt: In an outdoor courtyard with a textured stone wall, green potted plants, and out-of-focus trees in the background, two women stand on the left facing an older man seen from behind on the right. The blonde woman on the far left has shoulder-length blonde hair and wears a black leather double-breasted blazer with silver buttons over a white collared blouse, hoop earrings, and a delicate necklace. Beside her, the brown-haired woman has long wavy light brown hair in a half-up bun and wears a red and white striped sweater with a patterned red neck scarf and hoop earrings. The older man has short gray hair and a beard, wearing a dark coat over a patterned ascot. The brown-haired woman in the red striped sweater speaks first, holding a subtle knowing smile. Then the blonde woman in the black blazer begins to speak, becoming animated and extending her right hand toward the man to emphasize her point, breaking into a wide smile. The brown-haired woman shifts into a broad, amused grin. Throughout, the older man remains still, listening attentively to both women. Naturalistic, cinematic, pleasant tone.

Audio Prompt: In a quiet outdoor setting with a faint, low-level room tone, a young woman [Speaker A] speaks in a slightly exasperated tone: \"Please. He's always forgetting important stuff." A second young woman [Speaker B] responds in a light, teasing tone: \"Like our names." A soft, breathy chuckle is heard from [Speaker A]. [Speaker B] continues in a calm tone: \"This isn't anything new. Anything we need to be worried about."



Video Prompt: On a brightly lit residential porch with blurred green foliage in the background, an elderly man with short white hair swept back, a full white goatee and mustache, wearing a dark brown corduroy jacket over a dark plaid collared shirt with a blue and white patterned bandana at his neck, stands between two figures — a blurred shoulder on one side and a woman seen from behind with pulled-back blonde hair, glasses, and a light pinstriped shirt on the other. He speaks first with a serious, concerned expression, brow furrowed, leaning slightly forward. He then raises his right hand and emphatically taps his extended index finger against his right temple, his mouth opening wide in a sudden. His demeanor then shifts — he lowers his hand, looks down pensively, and when he looks back up, his face softens into a gentle, fond smile that turns into a quiet chuckle with a slight nod.

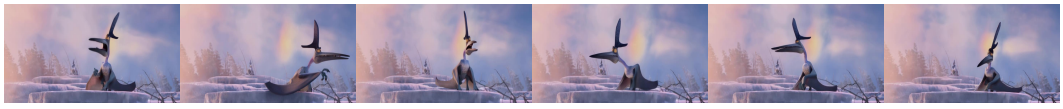
Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a mature man [Speaker A] speaks in a calm, conversational tone: \"You work with Ben?\" A mature woman [Speaker B] responds in a neutral, matter-of-fact tone: \"I do.\" The man continues, his voice slightly rising in pitch: \"So you're a cop?\" The woman replies, her tone measured and slightly corrective: \"Well, I'm a detective, right.\" The man responds with a drawn-out \"Ah,\" then says in a pleased, slightly relieved tone: \"Well, good. Hahaha, you know"



Video Prompt: In a hospital room with grey tiled walls, large windows letting in natural daylight, IV stands, and a blue monitor, a white female patient with long reddish-brown hair lies in a hospital bed on the right, covered by a blue patterned blanket, wearing a light-colored gown over scrubs. She grimaces in discomfort with her mouth open wide. In the adjacent bed, a young woman with shoulder-length blonde hair and bangs, wearing a white polka-dot hospital gown and a wristband, slowly sits up. She turns toward the distressed older woman, her expression shifting from concern to surprise as their eyes meet. To the left, an out-of-focus male nurse in green scrubs and a guard attend to her. Cool-toned palette of blues, greys, and whites contrasted by warm skin tones and reddish-brown hair. Realistic drama with a tense tone.

Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a mature woman [Speaker A] speaks in a panicked, high-pitched voice: \"They're gonna eat my face off." A mature woman [Speaker B] responds in a calm, steady tone: \"Oh my god. Okay, okay." A soft, melancholic piano melody begins to play in the background. [Speaker B] continues, his voice still anxious: \"Who are they" The woman [Speaker A] asks in a neutral, inquisitive tone: \"The flies" [Speaker B] replies in a strained voice: \"And how do you know they're gonna eat you"

Figure 13: Synthesized video-audio content involving multi-speakers. Please refer to the demo video for audio.



Video Prompt: On a flat-topped, blocky ice ledge in a prehistoric landscape of layered ice terraces and snow-covered ground, with sparse thin trees and a dramatic twilight sky of pink, purple, orange, and blue clouds behind, a large greyish-blue pterosaur with a purplish neck, white belly, long sharp-toothed snout, yellow eyes, tall dark curved crest, and large wings with warm orange-brown-white undersides stands with a wide-eyed, open-mouthed expression of shock. A small dark pterosaur stands behind the big one. The expression of the large greyish-blue pterosaur shifts through annoyance to a sly smile. It briefly spreads one wing wide, then settles into a thoughtful pose, resting its head on its clawed hand while gazing off-screen to the right. Its expression turns to concern, mouth opening slightly. Soft, diffused twilight lighting with no harsh shadows. Stylized 3D animation with warm pastel pinks and oranges in the sky contrasting cool blues and whites of the icy environment.

Audio Prompt: In a quiet outdoor setting with a faint, low-level room tone, a mature man [Speaker A] speaks in a calm, conversational tone: \"Yeah, well, uh, maybe. Oh, but first, we just need to check the area. There's plenty of folks in a very bad way after a storm like this, don't you know?".

Figure 14: Cartoon video-audio content. Please refer to the demo video for audio.



Video Prompt: At dusk in a desolate clearing beside a rustic log cabin with a thatched roof, shuttered windows, and barrels leaning against its side, under bare trees and a dim overcast sky, a bearded white man with short dark hair, wearing a loose brown shirt and matching pants, squats before a small crackling campfire within a stone ring on dry grass. To his right stands a slender white teenage boy with curly dark hair in an off-white Henley shirt and light trousers, holding a crumpled grey cloth. The man rises from his squat, reaches out with both hands, and takes the cloth from the boy. He turns toward the fire, steps forward, and bends down to drape the cloth over the burning logs. Smoke rises. The boy stands still, watching intently. The man then picks up a long wooden poker from the ground and pokes the smoldering bundle beneath the cloth.

Audio Prompt: A quiet outdoor dusk atmosphere with faint wind rustling dry grass. A small campfire crackles and pops within a stone ring. The soft rustle of cloth being handled. The crackling intensifies briefly as the cloth is draped over the fire, then dulls into a low, muffled smolder. A wooden poker scrapes against stone and prods the embers, stirring hissing smoke.



Video Prompt: At the water's edge facing the Manhattan skyline, an elderly man with white hair stands with his back to the camera. He wears a tan straw fedora, a dark zip-up bomber jacket over a maroon t-shirt, dark blue athletic track pants. He turns slightly and throws something small from his right hand toward the left; the object arcs out of frame before landing. He watches its trajectory off-screen. He then walks forward slowly along the water's edge, turning more toward the viewer with a thoughtful, calm expression as a seagull glides across the frame between him and the skyline.

Audio Prompt: In an outdoor waterfront setting with ambient waves lapping against the shore and a gentle breeze, a small object splashes into the water off-screen. A seagull cries as it glides past. Then a man [Speaker A] with a deep, gravelly voice speaks in a slow, deliberate, and slightly weary tone: /"So, you fed a little cheese to the cop, so what"



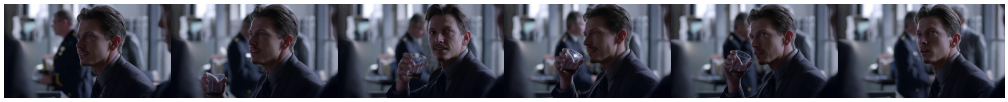
Video Prompt: Inside a stone-walled corridor with a paneled wooden door, a green-lit control panel on the wall, and a small vine plant growing nearby, under soft overhead lighting, five armored figures gather. On the right side, a clone in dark grey-black heavy armor with red accents and a backpack bearing a red emblem stands beside a taller clone in dark grey-black armor with a glowing orange-yellow eye slit. On the left side, a clone with grey-white armor and a yellow-orange visor, carrying a backpack with twin antennae, stands nearest to the door, operating the control panel. Behind him, the Commander — identifiable by his silver-faceplated helmet with yellow visor and grey armor with red shoulder markings — gestures and issues orders to the group. A lighter grey-white armored figure with a red visor stands behind the Commander. The two dark-armored clones on the right enter the doorway one after another. Then the twin-antennae backpack clone by the door steps through. The Commander follows next, striding through the doorway. Finally, the lighter grey-white armored figure behind the Commander trails him through the door last. 3D CGI animation style.

Audio Prompt: In a quiet indoor environment, a mature man [Speaker A] speaks in a calm, steady tone: /"You two, clear the upper levels." He [Speaker A] then continues in a neutral tone: /"Jack! We'll take the main floor and below. Droid, you're sticking with me.". A low, mechanical whirring sound begins and continues as the dialogue concludes.



Video Prompt: In a suburban park, a teenage boy and a young woman stand facing each other, initially heavily backlit as silhouettes against the low-hanging sun with a horizontal lens flare across the frame. The boy has short brown hair and wears a two-tone baseball t-shirt with mustard-yellow long sleeves and a grey torso with dark pants. The woman has wavy shoulder-length reddish-brown hair and wears a white textured lace dress. As the intense backlighting gradually softens, the boy's expression shifts from neutral to a slight smile as he speaks, then turns bashful — he looks down, brings his hands together.

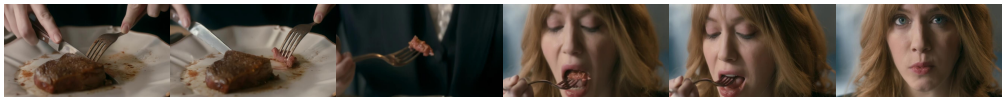
Audio Prompt: In a quiet outdoor setting with faint ambient sounds, a young man [Speaker A] speaks in a calm, friendly tone: "Well, happy birthday." A young woman [Speaker B] responds in a cheerful, appreciative tone: "Thanks."



Video Prompt: In an indoor setting with cool lighting and blurred figures in the background, a man with short brown hair swept back and light stubble sits in profile on the right frame, wearing a charcoal grey suit jacket, matching shirt, and a dark patterned tie. He looks toward the right with a knowing smirk. He slowly raises his right hand, bringing a faceted rocks glass containing dark amber liquid to chest height, holding it steady between himself and the unseen person. He then lowers the glass out of view, his expression softening as his mouth parts, transitioning from a tense stillness to active engagement as if about to speak. Cinematic realism.

Audio Prompt: In a quiet indoor setting, a mature man [Speaker A] speaks in a low, gravelly, and serious tone: "Show that sheet on TV." A soft, melancholic piano melody begins to play in the background. After a brief pause, the same man [Speaker A] speaks again in a steady, neutral tone: "About it, Walker."

Figure 15: Complex scene results (1/5). Please refer to the demo video for audio.



Video Prompt: In a dimly lit interior, a close-up shows hands using a knife and fork to slice through a medium-rare steak on a white square plate. After cutting off a piece, the camera tilts upward to reveal a Caucasian woman with long wavy reddish-blonde hair, blue eyes, and fair skin, wearing dark clothing over a collared shirt. She lifts the piece of steak with the fork, brings it to her mouth, and chews slowly. After swallowing, her gaze lowers briefly, then she raises her head and stares intensely forward.

Audio Prompt: A knife sawing through steak with a soft, wet slicing sound against the plate. A fork scrapes briefly. Quiet, slow chewing follows. After a pause, a single melancholic piano note rings out.



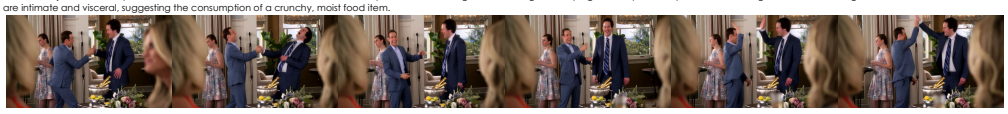
Video Prompt: The shot opens focused on a coastal village — orange-tiled roofs, a blue bay with moored white yachts — framed by out-of-focus green foliage and a light fabric awning in the foreground. The focus shifts as a middle-aged Caucasian woman with dark brown wavy hair in a low ponytail enters from the left, wearing small gold drop earrings, a white scoop-neck t-shirt with a small red heart graphic, and a woven strap bag over her shoulder. She approaches a large bundle of dried red chili peppers and green herbs hanging from strings, reaches out with both hands to gently touch them, then brings her face close. She closes her eyes and takes a slow, deliberate breath to inhale their scent. After a moment she pulls back, turns to the left, and walks down a slight incline away from the camera into an outdoor market path, passing other indistinct figures.

Audio Prompt: A gentle, melodic instrumental piece plays, featuring a prominent flute-like instrument and soft, rhythmic accompaniment, creating a calm and slightly melancholic atmosphere. The music is layered over the ambient sounds of a public outdoor space, including the indistinct murmur of people talking.



Video Prompt: In front of a solid black background, a young East Asian man with short black hair and a black t-shirt sits behind a light wood cutting board loaded with a cheeseburger, fried chicken pieces, pizza slices, french fries, and two bowls — one of ketchup, one of yellow mustard sauce. He looks at the camera, then lowers his gaze toward the food in contemplation. He ears a black-glove and takes a red Coca-Cola can. He takes the can, raises it, and drinks from the top.

Audio Prompt: In a close-up environment, the air is filled with the loud, wet, and rhythmic sounds of vigorous chewing and slurping, accompanied by distinct crunching noises as something hard is bitten into. The sounds are intimate and visceral, suggesting the consumption of a crunchy, moist food item.



Video Prompt: At a indoor party with large windows showing greenery outside, beige curtains, a silver champagne ice bucket, bowls of fruit, and pink roses on a low table, two blurred female figures frame the foreground on the left and right edges of the shot. Between them in focus, two middle-aged white men stand near a serving table. The man on the left in a blue-grey plaid suit jacket over a light shirt leans forward energetically and grabs the right hand of the man on the right in a dark navy suit, white shirt, and blue-black striped tie. They clasp hands and the plaid suit man laughs boisterously, mouth wide open. The navy suit man joins in, throwing his head back and swaying with uncontrollable laughter. To the left, a white woman with brown hair in an updo, wearing a sleeveless floral A-line dress, stands there, holding a drink in her left hand. The plaid suit man breaks free from the handshake and points excitedly at something off-screen, then both men raise their hands for a high-five.

Audio Prompt: In a lively indoor setting with a low murmur of background conversation, a man [Speaker A] lets out a sudden, loud, and playful "Ah!" followed by hearty laughter. The second man [Speaker B], with a deep, amused voice, says, "This guy's great." A first man [Speaker A], speaking in a casual, friendly tone, replies, "All time, man." The second man [Speaker B] then says, "I bought," as a sharp, distinct slap sound is heard.



Video Prompt: On a paved walkway outside a modern glass-fronted building with white pillars, a young Caucasian woman with long light-brown hair in one braid, wearing a floral sleeveless dress, performs on an orange electric violin with intense concentration. Pedestrians stroll past, with a man in a blue LA baseball cap watching intently. The violinist moves along the sidewalk as the handheld camera tracks her. Observational documentary style, enthusiastic tone.

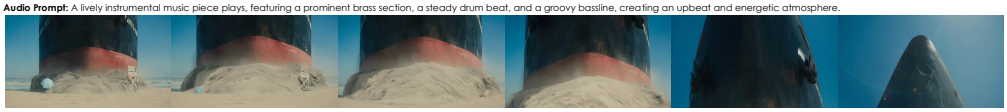
Audio Prompt: A melancholic and expressive instrumental piece plays, led by a solo violin with a rich, emotive tone. The violin's performance is accompanied by a gentle, rhythmic piano melody and a soft, steady bass line, creating a poignant and reflective atmosphere. The music unfolds with a slow, deliberate tempo, evoking a sense of longing and introspection.

Figure 16: Complex scene results (2/5). Please refer to the demo video for audio.



Video Prompt: In a warmly lit restaurant, a close-up high-angle shot centers on a bright orange ceramic hot pot filled with clear, steaming broth atop a dark wood table. The camera moves to a large black tray laden with shells, whole hard-boiled eggs, firm tofu pieces, and other vegetables. A pair of black chopsticks enters the frame, delicately picks up a orange shell, and lifts it toward the simmering broth as steam rises prominently. The chopsticks place the shell into the boiling liquid among the meaty chunks already cooking in the soup.

Audio Prompt: A lively instrumental music piece plays, featuring a prominent brass section, a steady drum beat, and a groovy bassline, creating an upbeat and energetic atmosphere.



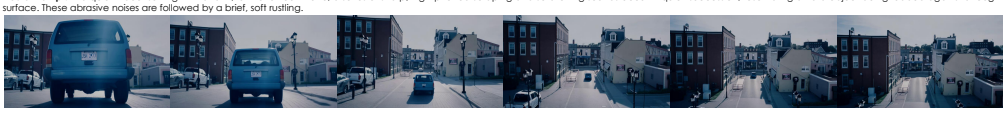
Video Prompt: On a bright sunlit beach, the colossal bow of a weathered black-hulled cargo ship with a faded red waterline plows through the sandy shore, pushing a massive wave of sand forward. The sand engulfs a lone light-blue patterned beach umbrella on the left and a wooden lifeguard chair on the right with a lifebuoy attached. As the ship advances, the camera slowly tilts upward along the sheer dark metal hull — revealing rust stains and scratches — rising from the chaotic destruction below to the pointed tip of the bow against a vast clear cyan sky, with the sun casting a subtle lens flare.

Audio Prompt: Set in a open environment with a low, continuous rumble of heavy machinery or a large vehicle, a deep, sustained mechanical hum fills the soundscape. The sound of a powerful engine idling is prominent layered with the subtle, rhythmic clanking of metal components. A sharp, metallic scraping noise cuts through the ambient drone, followed by a brief, high-pitched squeal. The engine sound then shifts, increasing in pitch and intensity. The entire scene is dominated by industrial, mechanical sounds, with no speech or music present.



Video Prompt: From a high-angle top-down view over a dark-brown wooden workspace, a pair of light-skinned male hands carve a small pale yellowish block shaped like part of a dog's body. His left hand grips the block firmly while his right hand holds a black-handled metal chisel, shaving away thin layers of wood in long curling shavings that accumulate around him. He turns the piece to carve from different angles, revealing a simple sketch drawn into its surface. In the blurred background, several finished painted Basset Hound figurines stand on a wooden board.

Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a series of sharp, high-pitched scraping and scratching sounds occur in quick succession, resembling a hard object being rubbed against a rough surface. These abrasive noises are followed by a brief, soft rustling.



Video Prompt: From behind a stationary light blue car with white plates, the vehicle begins moving slowly forward along a brick-paved downtown street. Multi-story red-brown and beige commercial buildings line both sides, adorned with ornate black gas lamps; The sky is overcast. The car proceeds straight ahead, eventually disappearing toward a distant intersection with traffic lights. The shot transitions from a close-up to a wide establishing shot via a slow zoom-out. High-angle, centered composition initially focused on the vehicle before revealing the full street context with strong leading lines toward the horizon.

Audio Prompt: A low, ominous, and suspenseful instrumental music track plays throughout, featuring deep, pulsating synthesized bass tones and a slow, building rhythmic pattern that creates a tense and dramatic atmosphere. The music swells in intensity, with layered electronic textures, but no speech or other distinct sounds are present.



Video Prompt: In a multi-level corporate office high above a city skyline, a Black man with short black hair, wearing a dark blue suit jacket over a white collared shirt unbuttoned at the top and dark trousers, steps off an staircase bordered by angular wooden partitions onto the grey-carpeted main floor. He glances up briefly, then walks with a confident stride toward the right. The camera pans right, tracking him as he passes a woman with long brown hair at a desk. She notices him, stands up, and speaks to him as he walks by. He continues past open-plan areas toward glass-walled offices overlooking a dense urban skyline. He approaches one office where a bald white man holding a white cup stands talking to another man seated behind a large desk, while a woman in a blazer and a women's tailored pant stands nearby with arms crossed.

Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a woman [Speaker A] speaks in a calm, slightly urgent tone: "He's in a meeting." A man [Speaker B], whose voice is deeper and more assertive, responds: "Tll tell you and your lawyer at the same time. Your attempted purchase of Enzy Novo has hit a snag."

Figure 17: Complex scene results (3/5). Please refer to the demo video for audio.



Video Prompt: At water level on the open ocean, a small translucent jellyfish-like creature with an iridescent, dome-shaped float drifts in the foreground, bobbing with deep blue undulating waves. The setting sun hangs low on the horizon behind thin clouds, casting a warm yellow-orange gradient across the sky and a golden path over the water's surface. The camera slowly tilts up from the drifting creature, revealing the expansive rippling sea. A dark silhouette resembling a bird sweeps swiftly across the sky above the water. The camera then tilts back down to settle on the iridescent float against the sunset backdrop. Golden-hour backlighting illuminates wave edges and the translucent float, creating dramatic contrast. Deep cool blues of the ocean contrast with warm saturated yellows and oranges of the sunset. Naturalistic cinematic style, serene and meditative tone.

Audio Prompt: A dramatic orchestral score plays, featuring sweeping strings and a deep, resonant bass that creates a sense of suspense and grandeur. The music swells and then recedes slightly, maintaining a tense, cinematic atmosphere. Over this, the distinct squawking of a bird—likely a seagull—can be heard, its calls sharp and insistent. The music then fades out, leaving only the bird's squawks for a moment.



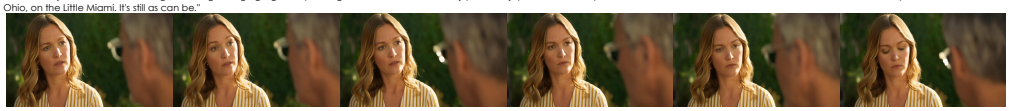
Video Prompt: In the driver seat of a moving car, a fair-skinned white woman with long wavy reddish-blonde hair blowing slightly in the wind wearing a grey heathered long-sleeved shirt. She stares forward with a pensive, vacant expression. As blurred suburban scenery passes the open window, she slowly brings her left hand to her face, pressing fingers to her forehead then running them back through her hair. After a moment, she leans her elbow on the open window frame, propping her head against the side of her hand at her left temple. Her eyes open but holding a look of deep concern and exhaustion as the drive continues. Cinematic realism, somber tone.

Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a mature woman [Speaker A] speaks through a phone speaker in a tinny, slightly distorted voice with a shaky, distressed tone: /"Hi honey, how are you?" A young woman [Speaker B] responds in a calm, concerned tone: /"Hi honey, can I um... can I talk to Corey?" [Speaker A] replies in a worried, urgent tone: /"Sweetheart, what is it? Are you okay?"



Video Prompt: On a wide, murky-green river lined with dense green foliage and leafy trees forming a natural canopy, an older man with short graying hair and glasses sits in a small boat. He wears a beige t-shirt under a yellow and black life jacket with a microphone clipped near him, holding a paddle. He looks at the camera and talks expressively, turns slightly away, then turns back continuing his dialogue. He gestures by pointing forward and then outward across the water's surface.

Audio Prompt: In an outdoor setting with the gentle gurgling and splashing of water, a mature man [Speaker A] speaks in a steady, conversational tone: /"on the waters. Now, what do I want to buy? Well, here I am in Ohio, on the Little Miami. It's still as can be."



Video Prompt: Outdoors against soft-focus lush green foliage, an over-the-shoulder close-up frames a Caucasian woman with shoulder-length wavy blonde hair, light eyes, wearing a mustard-yellow and white striped collared blouse and small gold hoop earrings on the left third of the frame. To her right, the blurred profile of an older man with graying hair and light-framed sunglasses is visible. Her lips are slightly pursed as if finishing a point. Her expression then transitions — she closes her mouth, her brow furrows slightly, and her gaze becomes fixed and intense, conveying deep concentration as she listens to the speech of the man.

Audio Prompt: In a quiet outdoors setting with a faint, low-level room tone, a mature woman [Speaker A] speaks in a calm, explanatory tone: /"We're going to do when you took it". A mature man [Speaker B] responses in a excited tone: /"Look, I knew it didn't belong there, so yes, of course, finding its rightful owner, it would be the next appropriate step."



Video Prompt: In an outdoor residential setting with a two-story brick house in the distance, green grass with scattered dry leaves, tall pine trees, and a lone wooden bench on the lawn, a white man with dark brown hair in a tight ponytail and a trimmed full beard stands wearing a teal-blue denim button-down shirt with sleeves rolled to his elbows. In the foreground, two out-of-focus figures frame the shot — the outermost showing only light blonde long hair at the edge of the frame, and the second revealing a short blonde-haired head and shoulder. He looks seriously at them, then begins to speak, becoming more animated. He raises both hands in front of his chest — his right hand clenched into a fist and his left hand open with palm facing up — and repeatedly strikes his right fist into his left palm to emphasize his points. His expression shifts from serious concentration to passionate explanation. Cinematic realism.

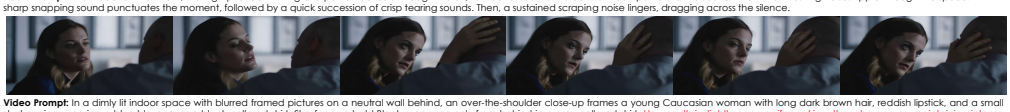
Audio Prompt: In a quiet outdoor setting with a faint ambient breeze, a mature man [Speaker A] speaks in an agitated, urgent, and conspiratorial tone: /"Who they really work for, Aaron. They launder drug money." His hands clap together sharply for emphasis. /"Drug money, Aaron," — another firm hand slap — /"for the Navarro drug cartel." Rhythmic tapping sounds punctuate his words as he gestures intensely throughout.

Figure 18: Complex scene results (4/5). Please refer to the demo video for audio.



Video Prompt: From a high-angle overhead view on a polished stainless steel countertop, a pair of light-skinned male hands work methodically. One hand holds a large wooden spatula, stirring a creamy off-white sauce bubbling gently in a silver frying pan atop a red and grey portable induction stove. The other hand operates a black electric spice grinder grinding black pepper directly onto the hot sauce. He stirs to incorporate the spices, pausing after grinding doses. Surrounding items include metal tongs, and two glass jars of orange and red granulated seasonings.

Audio Prompt: In a indoor environment, the high-pitched whirring of a power tool cuts through the air, its mechanical hum a constant companion to the scene. Intermittent rustling noises ripple through the space. A sharp snapping sound punctuates the moment, followed by a quick succession of crisp tearing sounds. Then, a sustained scraping noise lingers, dragging across the silence.



Video Prompt: In a dimly lit indoor space with blurred framed pictures on a neutral wall behind, an over-the-shoulder close-up frames a young Caucasian woman with long dark brown hair, reddish lipstick, and a small stud earring, wearing a blue blazer over a black collared shirt. She faces a bald Black man seen only from behind in a grey collared shirt. Her mouth is slightly open as if speaking, then she pauses, maintaining intense eye contact with deep concentration. She slowly raises her right hand, palm inward, and gently places it against the side of his head, cupping it near his ear and neck. Her gaze remains fixed on him, lips slightly parted.

Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a mature woman [Speaker A] speaks in a soft, intimate, and breathy tone: /"I can touch you." A gentle, slow brushing sound of skin against skin accompanies her words. She continues in the same gentle manner: /"Do you like that when I touch you?" The soft, rhythmic sound of her hand lightly caressing continues underneath her voice.



Video Prompt: Inside a brightly lit boxing ring before a dimly lit crowd, a muscular Black man with a mohawk and beard, wearing white trunks and red gloves, faces off against a muscular white man with dark curly hair and a cut on his forehead, wearing red gloves and blue-red-white striped shorts. The Black boxer attacks with a rapid-fire combination of punches. His opponent ducks and blocks the flurry. They circle each other exchanging jabs, hooks, and crosses, sweat glistening under harsh overhead spotlights. After an aggressive sequence, the white boxer lands a powerful right hook to the head. A referee in a light blue shirt steps between them, placing hands on their shoulders to break up the clinch. Gritty, realistic 1980s action film style with film grain and motion blur.

Audio Prompt: Set in a chaotic, high-energy environment with a driving, rhythmic hip-hop beat playing in the background, a series of sharp, percussive impact sounds and grunts suggest a physical altercation. The crowd [Speaker A], speaking in an aggressive, confrontational tone, shouts: /"He's so bad! He's so bad! He's so bad!" His voice is strained and forceful. Another part of crowd [Speaker B], also speaking in an aggressive tone, retorts: /"Why he so bad? Yeah" A man [Speaker C] shouts: /"Enought" The background music and sound effects continue throughout, creating a tense and combative atmosphere.



Video Prompt: On a cracked asphalt driveway in a residential backyard with a white-sided house and garage on the left, a tall evergreen tree, and a chain-link fence, a young man with short blondish-brown hair in a black t-shirt and black athletic shorts with red and white stripes stands facing away from the camera. At the far end of the driveway, another young man in a camouflage baseball cap, brown and green plaid flannel shirt, and light-wash denim pants waits. The player in black passes the orange basketball to the flannel-shirted player, who catches it and passes it back. The flannel-shirted player then moves in to steal the ball, and the player in black plants his feet and holds a stationary defensive stance, shielding the ball with his body.

Audio Prompt: Set in an outdoor environment with the faint, continuous chirping of birds in the background, a young man [Speaker A] speaks in a calm, neutral tone: /"here." A soft thud is heard, followed by the sound of a ball being dribbled on a hard surface. Another young man [Speaker B] speaks in a steady, neutral tone: /"Alright." The dribbling sound continues, rhythmic and clear, as the ball bounces on the ground.



Video Prompt: In an indoor studio with a white backdrop curtain on one side, a black-painted wall on the other, and shelving with colorful props, a middle-aged Caucasian man with short brown hair sits at a light-yellow table wearing a grey t-shirt with a red logo. In front of him sit a cardboard box. He holds one blue LED yo-yos, drops it. As the yo-yo spins through the air, a second person's hand enters the foreground and catches it — the yo-yo now glowing with purple lights. The man resumes talking and gestures toward the camera. The second man then throws the yo-yo back to the first man.

Audio Prompt: In a workshop environment, the high-pitched whirring of a power tool cuts through the air. A sharp snapping sound punctuates the moment, followed by a quick succession of crisp tearing sounds. A mature man [Speaker A] speaks in a instructional tone: /"Oh, that was, that was angled back at me. So I'm gonna angle it forward just a little bit so it'll go to him. There we go. He's gonna toss it back. And you can't be hesitant when you put your hand"



Video Prompt: In an indoor martial arts dojo with red and black floor matting, black acoustic-paneled walls bearing large circular white dragon emblems, and a red ceiling with fluorescent tube lights, a Caucasian bold man shaved on the sides, blue eyes. He wears a dark grey t-shirt with a small microphone clipped on. He also wears a black watch. He speaks directly to the camera with expressive gestures, then demonstrates a series of punches — starting with a jab, transitioning to an uppercut and a hook.

Audio Prompt: In a quiet indoor setting with a faint, low-level room tone, a mature man [Speaker A] speaks in a steady, instructional tone: \"When I throw that punch, I don't want you to hook punch all the way, all the way across your body. I want you to throw that hook punch and it goes boom when it hits like."

Figure 19: Complex scene results (5/5). Please refer to the demo video for audio.

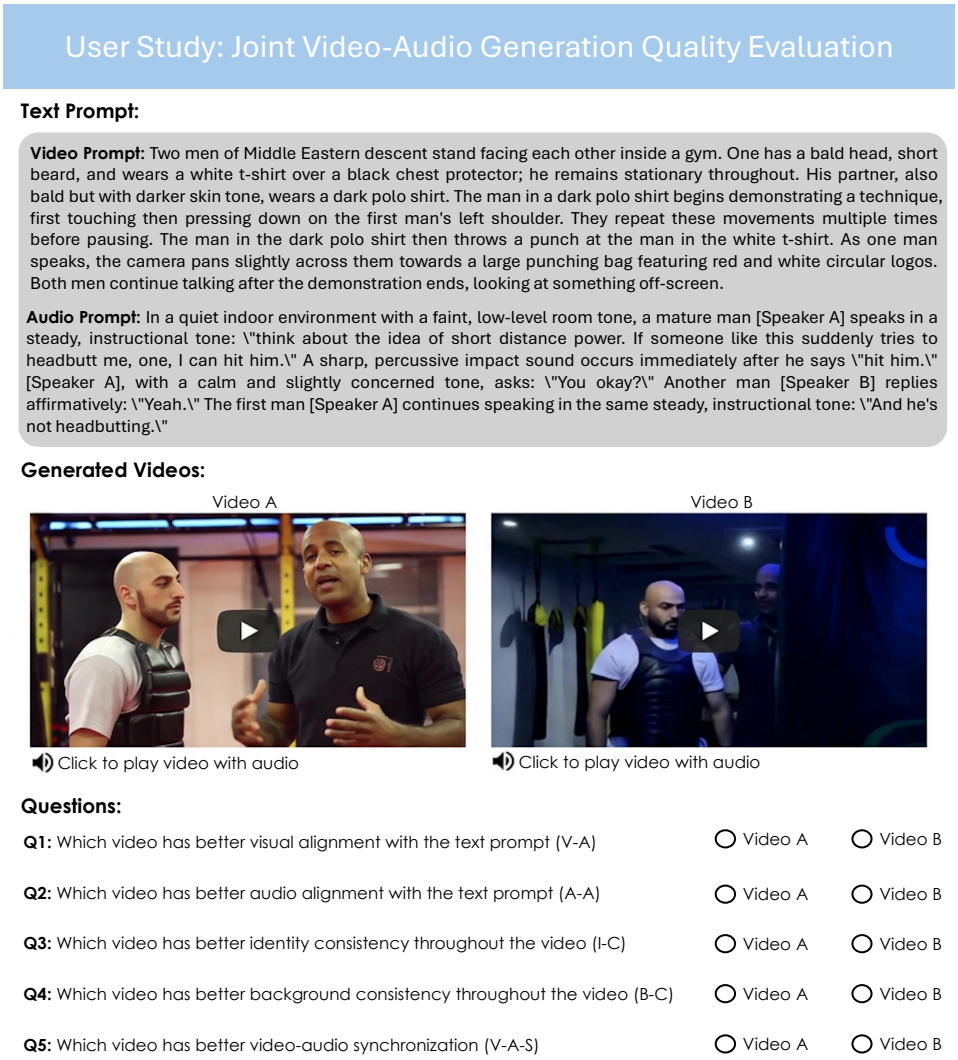


Figure 20: The user study screenshot.



Figure 21: One failure case of our Baton.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to the Introduction Section and Method Section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Sec. A.17 in the Appendices.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please refer to the Method Section and Experiments Section.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to the Method Section and Experiments Section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code and data are currently undergoing legal review. We will release them publicly once the review is complete. The details of the training and testing datasets are described in Sec.A.5 in the Appendices. The reproducibility details of our method are depicted in the Method Section and Experiment Section in the main paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Sec.A.6 in the Appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Sec. 4.4 in the main paper and Sec.A.6 in Appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Please refer to Sec. A.16 and Sec. A.15 in the Appendices.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Sec. A.16 in the Appendices.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Please refer to Sec. A.16 in the Appendices.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please refer to Sec. A.5 and Sec. A.16 in the Appendices.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Please refer to Sec. A.5 in the Appendices.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Please refer to Sec. A.11 in the Appendices.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Please refer to Sec. A.11 in the Appendices.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We utilize an LLM (Qwen3) to initialize our model backbone. Please refer to the Method Section and Experiment Section in the main paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.