Baton: Explicit Semantic Blueprints for Joint Video-Audio Generation

Shuyuan Tu[1,2,*,‡], 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] Fudan University, [2] Tencent Hunyuan

Tencent Hunyuan

Institute of Trustworthy Embodied AI
Fudan University

Project & Demo

NEURAL INFORMATION PROCESSING SYSTEMS

Overview



Prompt: "A young man in a brown hooded coat lies prone, covered in sand, as distant explosions erupt behind scattered soldiers. He flinches in terror and clutches his head while successive blasts draw closer, sending towering columns of sand into the air. A harsh, aggressive soundscape with deep rumbles and piercing screeches builds as the bombardment rapidly approaches his position."



Prompt: "The instructor on the right demonstrates by delivering a quick punch to the left man's upper body, producing a sharp impact sound, after which the camera shifts focus to the man on the left. The instructor then asks, "You okay", the other replies "Yeah," and the instructor continues his explanation." **(a)**

Motivation

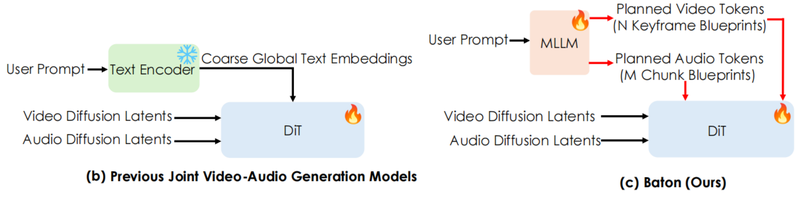
(1) Existing joint video-audio diffusion models rely on coarse global text embeddings that lack fine-grained temporal semantics. Consequently, they struggle with multi-stage actions, human-object interactions, and multi-speaker dialogues, often causing visual distortions and audio-visual misalignment.
(2) Without a shared semantic plan, video and audio denoising trajectories may diverge under complex prompts, leading to unstable generation. This motivates **temporally structured semantic blueprints that explicitly coordinate both modalities**.

Contributions

(1) We propose **Baton**, a framework that **disentangles semantic planning from joint video-audio synthesis**. By establishing shared temporal blueprints before denoising, Baton coordinates both generation trajectories for stable and synchronized generation.
(2) We introduce **VA-Planner**, an MLLM with dual semantic alignment towers. Learnable queries and bidirectional cross-modal attention fuse **visual and auditory semantics to produce mutually aligned planned tokens**, providing fine-grained, frame-level guidance beyond global text embeddings.
(3) We propose **Relative Semantic RoPE (RS-RoPE)** to **align planned tokens and diffusion latents in a shared spatiotemporal coordinate system**, enabling position-aware cross-attention while preserving structured semantic guidance during denoising.

Implementation

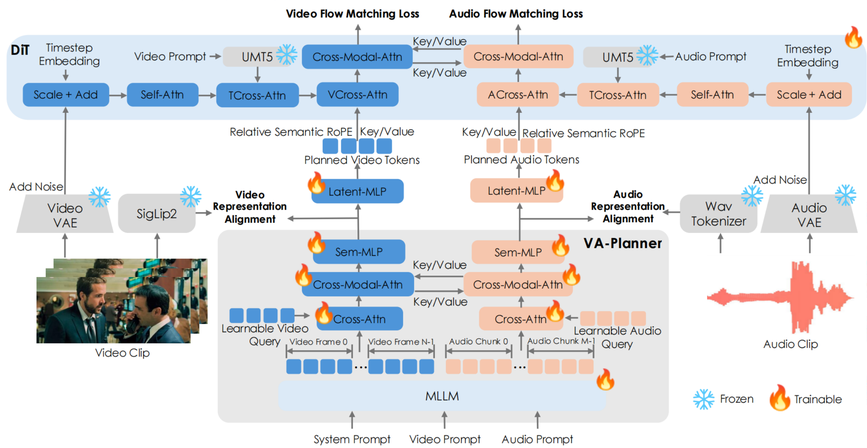
Qwen3-8B VA-Planner + **Ovi** DiT backbone | 1.5M training clip
Video plans: 6-FPS keyframes **Audio plans:** 1-second chunks



(b) Previous Joint Video-Audio Generation Models **(c) Baton (Ours)**

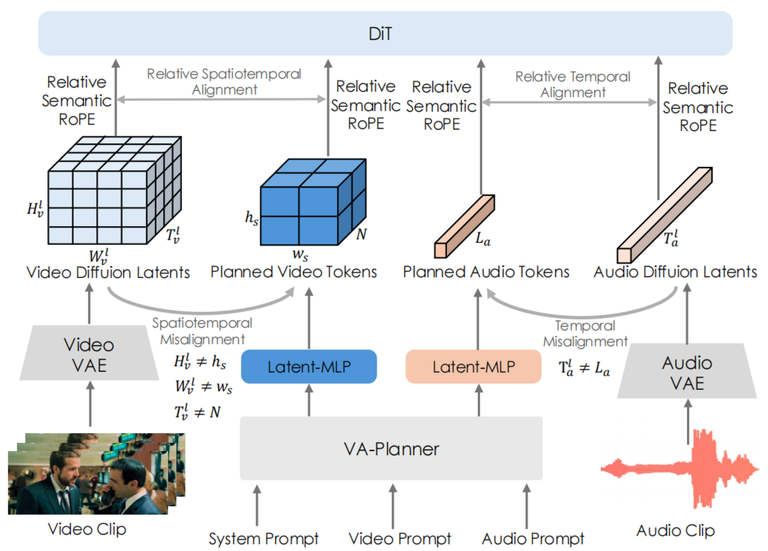
Methods

Baton: VA-Planner + Joint DiT



The MLLM predicts **video keyframe plans** and **audio chunk plans**; dual semantic towers mutually align them before cross-attention-based generation.

Relative Semantic RoPE (RS-RoPE)



Video: 3D time–height–width alignment. Audio: 1D temporal alignment.

Three-Stage Training

STAGE 1
Planner pretrain
Regress SigLip2 / WavTokenizer features.

STAGE 2
DiT adaptation
Flow matching with ground-truth features.

STAGE 3
Joint fine-tune
Freeze planner; condition DiT on predicted plans.

Results

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 (Low et al., 2025)	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 (Liu et al., 2025a)	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 (Zhang et al., 2025)	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 (HaCohen et al., 2026)	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 (Team et al., 2026)	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**



Video Prompt: On a sunny suburban backyard with green lawn and tall hedges, a woman in a ribbed sweater and black skirt rallies a shuttlecock with a boy across a badminton net. He jogs off-screen to fetch it; she turns toward camera, striding forward with arms raised in playful triumph. A second boy charges in from behind, tackling her into wrestling on the grass — she lifts him, spins him around, both laughing joyously. Handheld tracking shot follows the action, shifting from eye-level to low-angle during the lift.

Video Prompt: On a residential street corner, a young Asian boy in bright blue shorts stands holding a brown Spalding basketball in one hand and a yellow-orange ball in the other. The camera slowly orbits from behind him onto a concrete patio beside a dense green hedge. He drops into a low stance and begins simultaneously dribbling both balls side by side, bouncing them in rhythm as he moves forward step by step along the road. Medium close-up widening to a tracking shot, eye-level shifting to slightly high-angle.

Audio Prompt: A fast-paced electronic dance music track with a driving beat and synthesized melodies plays throughout the clip. A boy [Speaker A] shouts excitedly in a energetic voice: \" Oh no! Ten points! I'm scared! She's the winner!" A girl [Speaker B] shouts back with equal excitement: \"We're the winners!"

Audio Prompt: Set in an outdoor environment, a young boy [Speaker A] speaks in a clear, instructional tone: \"This is two ball basketball drill.", immediately after he finishes speaking, the rhythmic, percussive sound of a basketball being dribbled on a hard surface begins and continues for the rest of the clip.



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.

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.

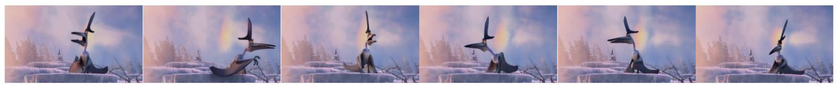
Audio Prompt: In a festive indoor setting with a low-level room tone, a mature man [Speaker A] speaks in a cheerful, enthusiastic tone: \"So we thought, might be a good idea to gather everyone together, and wish you a merry Christmas!" As he finishes, a burst of upbeat, orchestral holiday music swells, 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.

Audio Prompt: In a quiet 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.



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 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?",