# Video = World + Event Stream

**Wan Team, Alibaba Group**

See Contributions and Acknowledgements for the full author list.

## Abstract

We present **Wan-Streamer v0.3**, which reframes our native-streaming interaction model under a single organizing view: **a video is a world plus an event stream**. The *world* is the persistent context in which a video unfolds, including the environment, scene, subjects, ambient acoustic conditions, voice characteristics, and other relatively stable conditions. The *event stream* is everything that changes over time within that world, including scene or environmental changes, subject behavior, speech, and other sounds. This yields a **general-purpose pretraining task** over large amounts of real video: given a world and incoming input, predict how the world moves, changes, and responds in real time. The resulting competence can be specialized to a broad family of real-time downstream tasks. We instantiate it on **real-time full-duplex audio-visual interaction**, where the event stream is the agent's speech together with free-form behavior. Functionally, the model's multimodal understanding process is vision-language-action-like: it maps multimodal user input to language-form speech and behavior actions. Wan-Streamer v0.3 preserves the v0.2 operating point: **640×368** video at **25 FPS**, a **160 ms** streaming unit, approximately **200 ms** model-side response latency, and approximately **550 ms** total interaction latency under a 350 ms bidirectional network budget.

**Website:** https://wan-streamer.com/

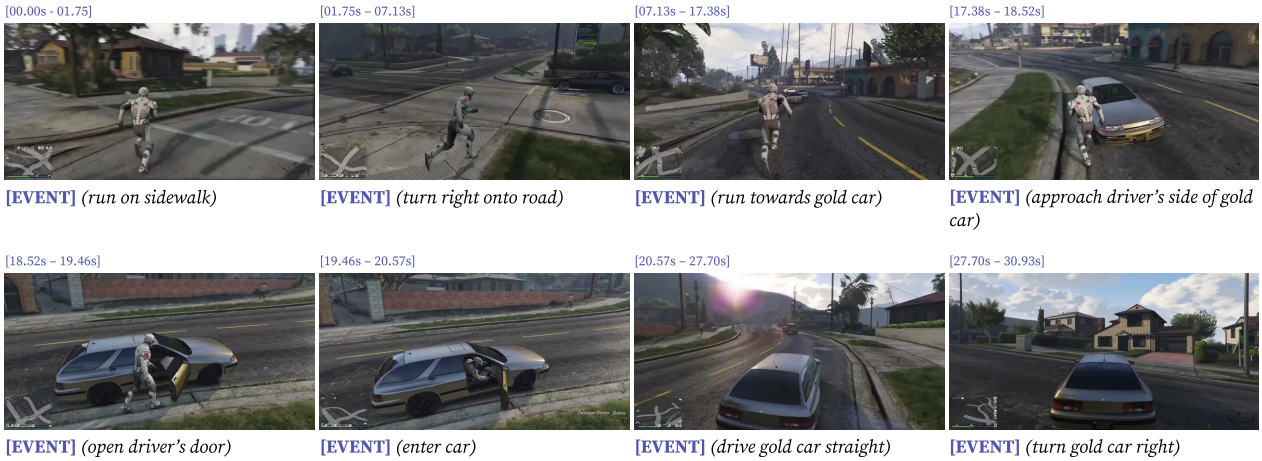
## 1 Introduction

Wan-Streamer v0.1 [1] established a native-streaming, end-to-end formulation for real-time full-duplex audio-visual interaction: user and agent text, audio, and video live on one causal timeline and are modeled by a single Transformer, so perception, response timing, speaking, visible listening, and synchronized video are learned as one behavior rather than assembled from cascaded modules. Wan-Streamer v0.2 [2] kept that formulation fixed and raised the interactive stream from 192×336 to 640×368 at the same ∼200 ms model-side latency, moving the expensive latent generation into a Ulysses-style context-parallel performer [3].

Both versions treat one target application—an agent that talks and reacts—as the training objective. Wan-Streamer v0.3 steps back and asks what the underlying competence is. Our answer is a simple decomposition: **any video can be read as a world plus an event stream**. The *world* is the persistent context in which a video unfolds: its environment and scene, visual style, subjects and their identities and appearances, ambient acoustic conditions, voice characteristics, and other relatively stable conditions. The *event stream* is everything that changes over time within that context: scene or environmental changes, subject behavior, camera motion, speech, and other sounds. A recorded video is one realization of an event stream unrolling inside a given world; an interactive experience is the same object with the event stream produced online in response to input.

This decomposition defines a general-purpose pretraining task. We pretrain on large amounts of ordinary video to learn one thing: **given a world and an incoming input, how does the world stream forward—how does it move, change, and respond, unit by unit, in real time.** Because essentially all natural video is an instance of a world with an event stream, this objective is broad and scalable, and the competence it produces is world knowledge about how scenes plausibly evolve. That competence can then be specialized by varying the world context and the input that drives the stream, while retaining the same streaming response logic.

**[WORLD #1]** *A suburban neighborhood during the daytime. The **environment** features paved asphalt roads with yellow lane markings, sidewalks, and grassy verges ... The **audio** includes various in-game sounds such as the character's footsteps on pavement, the engine sounds of passing cars ... **@character1** is a humanoid robot with a sleek, white and grey armored suit. It has glowing red circular lights on its chest and shoulders ...*




**[WORLD #2]** *The **environment** features an outdoor poultry pen scene with a green awning above. The ground is dirt. There is a chicken coop ... The **audio** includes a female voice, speaking English, with a friendly and instructive tone, and you can hear the clucking of chickens and the crowing of roosters, as well as the sounds made by characters walking ... **@character1** is a middle-aged white woman with blond hair wears a brown wide-brimmed hat. She wears a black T-shirt...*


**Figure 1** Two examples of the general-purpose pretraining data: time-aligned events paired with a summary of the persistent world context, including the visual setting, acoustic conditions, and characters. Parentheses denote character behavior, while quotation marks enclose spoken words. For clarity, both examples contain only one character, so actor names are omitted from the event labels. Training supports multiple characters by associating each event with its corresponding character, or with an explicit no-character marker for character-independent events.

We instantiate the pretrained world-event model on the same downstream task as v0.1/v0.2—real-time full-duplex audio-visual interaction—but widen what the agent can express. Here the *world* configures the scene, character, ambient sound, and voice; the *event stream* is the agent's response, driven by the user's streaming text, audio, and video. The key change from v0.2 is that the agent's event stream now carries **free-form behavior** alongside speech. We adopt a role-play chat format in which the model's language/state stream interleaves spoken words with parenthesized behavior directives, e.g. "*(frowning slightly)* What did you say?" or "*(picks up the mug and glances toward the window)* Sure, one moment." The parenthesized part can be any behavior describable in language, so the agent is not limited to a small fixed action vocabulary

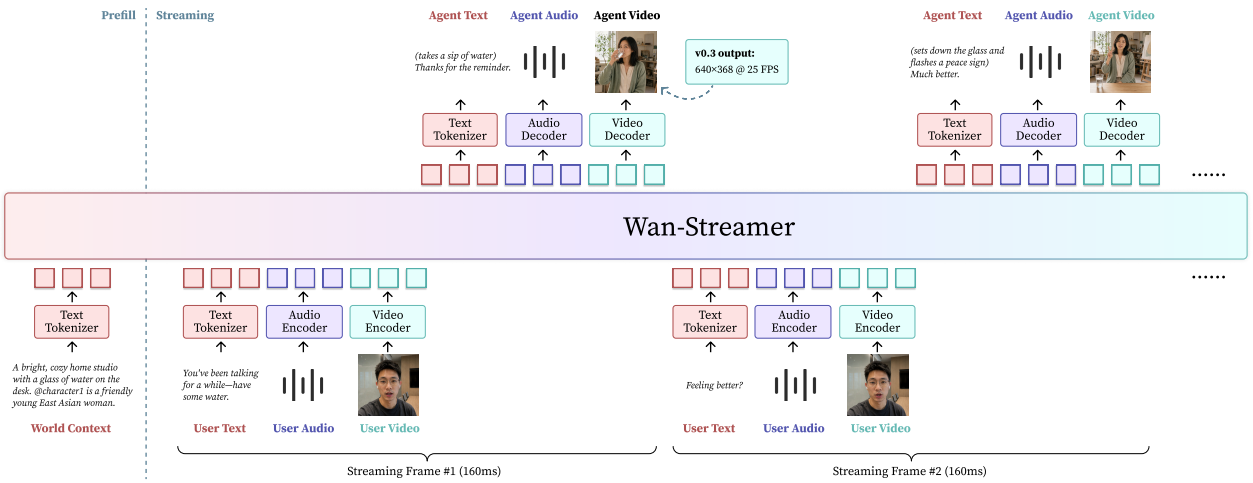



**Figure 2** The world context is tokenized and prefilled once before streaming. Language, audio, and video inputs and outputs then share a causal timeline coordinated by block-causal attention. The model maps this world and streaming multimodal user input to language-form speech and behavior actions. These predicted tokens condition synchronized audio-video generation. In the agent text stream, phrases inside parentheses denote free-form behavior, while text outside parentheses is spoken by the agent.

(such as the `w`/`a`/`s`/`d` moves used in roaming) but can perform open-vocabulary actions grounded in the world. This interleaved stream conditions audio-video generation exactly where the language stream did before, so no new latency-critical path is introduced.

Because the modeling contract and the deployment topology are unchanged, v0.3 preserves the v0.2 operating point: 640×368 video at 25 FPS, a 160 ms streaming unit, approximately 200 ms model-side response latency, and approximately 550 ms total interaction latency with a 350 ms bidirectional network budget. Figure 1 shows the general pretraining representation before Figure 2 shows its interaction instantiation. The rest of the paper formalizes the decomposition and objective, describes the downstream behavior interface, compares versions, and reports experiments. The serving topology is inherited unchanged from v0.2, so we do not restate it here.

Our contributions are:

- We reframe native-streaming generation as a **world + event stream** decomposition that separates persistent world context from time-varying events, yielding a general-purpose pretraining objective that transfers across real-time tasks.
- We instantiate this model on real-time full-duplex audio-visual interaction, where the model's multimodal understanding is VLA-like, mapping streaming multimodal input to speech and **open-vocabulary behavior actions expressed in natural language**. These language-form outputs condition synchronized audio-video generation.
- We show that this expansion keeps the v0.1/v0.2 modeling contract and the v0.2 serving topology intact, preserving 640×368 at 25 FPS with approximately 200 ms model-side and approximately 550 ms total interaction latency.

# 2 World-Event Decomposition

## 2.1 Formalization

We read a video as a persistent *world* together with a time-varying *event stream*. The world provides the context that remains relatively stable over the timescale of a clip or interaction—its visual setting and style, environment and layout, subjects and their identities and appearances, ambient acoustic conditions, and voice characteristics. The event stream contains the changes that unfold within that context, including subject behavior, camera motion, environmental change, speech, and other sound events.

For learning, we represent the world as a flexible structured record

$$W = \{w_j\}_{j \in \mathcal{S}(W)}, \tag{1}$$

where the schema $\mathcal{S}(W)$ is determined by the clip or task rather than fixed globally. In practice, fields commonly group visual context (medium or style, environment, layout), acoustic context (ambient sound and voice characteristics), and a set of character records (identity, appearance, persona, visibility), but fields may be absent or extended.

An event is a time-localized record

$$e_k = \big(\tau_k,\, c_k,\, d_k\big), \qquad c_k \in \mathcal{C}(W) \cup \{\varnothing\}, \tag{2}$$

where $\tau_k$ is its active interval, $d_k$ is a free-form description of the change, and $c_k$ associates the event with a character in $W$ or explicitly marks it as character-independent. The description may cover behavior, camera motion, environmental change, speech, or physical sound; overlapping changes can be represented by multiple records. A video is then a world with its event stream unrolled, and its observable audio-visual realization is produced by the causal decoders of v0.1 from the generated latents. This is not a new architecture: it is a reading of the same interleaved causal sequence in which durable conditioning is named $W$ and time-localized change is named $e_k$.

## 2.2 Streaming world-event factorization

The v0.1 formulation predicts the next response unit from the full causal history. Writing that response as the event unit $e_k$ and letting $x_{\le k}$ denote whatever input drives the stream (for pretraining, the observed past; for interaction, the user's streaming observations), the model factorizes as

$$p_\theta\big(e_{1:K} \mid W, x_{1:K}\big) = \prod_{k=1}^{K} p_\theta\big(e_k \mid W,\, x_{\le k},\, e_{<k}\big), \tag{3}$$

where $K$ is the number of streaming units. The v0.1 autoregressive factorization is the special case in which $W$ is folded into the history, $x$ is the user observation, and $e_k$ is the agent's text/audio/video response. As before, discrete parts of $e_k$ (language, and now behavior directives) are optimized with next-token prediction, while the continuous audio and video latents are generated with conditional flow matching under the same clean context, so speech, motion, appearance, and scene evolution are denoised as one coupled response and committed back into history for the next unit. Naming $W$ explicitly makes the persistent world context a first-class conditioning object: at inference it is provided once and the event stream is produced online.

## 2.3 Pretraining and downstream transfer

Equation (3) defines a general-purpose pretraining task over large-scale, diverse video corpora rather than one interaction domain. Each clip supplies an inferred world and time-aligned events, and the model predicts the next event and its audio-visual realization from causal history. This competence can in principle support many downstream interfaces, each defined by its world context and stream-driving input: world-model control and real-time first-person interaction use action or navigation input over a scene world [4–7]; embodied manipulation uses commands over a workspace world [8–10]; and our instantiation uses streaming user text, audio, and video over a scene-and-character world. This paper studies only the real-time full-duplex

audio-visual interaction specialization and leaves systematic adaptation and evaluation of the others to future work.

Figure 1 makes this supervision concrete: each example summarizes its persistent world context once using the information relevant to that clip, while natural-language events are aligned to the intervals in which they occur. Learning how these events unfold in context teaches the pretrained model world-event dynamics that can be reused when events are produced online at inference.

# 3 Real-time Interaction

## 3.1 Instantiating the world and the event stream

For real-time full-duplex audio-visual interaction, the persistent world context is instantiated with the settings needed by this task:

- **Scene setting**: the environment and layout the agent inhabits.
- **Character setting**: the agent's identity, appearance, and persona.
- **Ambient-sound setting**: the background acoustic field of the scene.
- **Voice timbre**: the agent's speaking voice.

The user's streaming text, audio, and video are the input $x_{\leq k}$ that drives the event stream. In v0.3, the agent response interleaves speech and free-form behavior records, both rendered into synchronized audio and video by the causal decoders.

This inference-time behavior stream is the controllable counterpart of the time-aligned behavior language used during pretraining in Figure 1: instead of being recovered from recorded video, the behavior event is produced or supplied online and the model streams the corresponding world evolution.

## 3.2 Free-form behavior in a role-play format

v0.1 and v0.2 focus on an agent that speaks and shows visible listening behavior—idle presence, gaze, nods, micro-expressions, lip-synchronized speech. v0.3 widens the behavior channel to **open-vocabulary actions described in natural language**. We adopt a role-play chat format in which the agent's language output interleaves spoken words with parenthesized behavior directives, for example:

> *(reaches into the grass and picks up a green leaf)* Look what I found.
> *(covers mouth in surprise)* I did not expect that.

The parenthesized directive can be any behavior expressible in language, so the agent is not confined to a small fixed action set (such as the `w`/`a`/`s`/`d` navigation moves used in roaming) but can perform situated, open-vocabulary actions that are grounded in the configured world—reaching for a nearby object, turning toward a sound, changing posture, or reacting expressively. Behavior directives and spoken words share one token stream, so their timing, ordering, and interleaving with speech are learned jointly rather than scheduled by an external controller.

## 3.3 Multimodal understanding and language-form actions

Functionally, the model's multimodal understanding process is vision-language-action-like: conditioned on the world, it continually maps the user's text, audio, and video to language-form speech and behavior actions. These actions form an interleaved speech-and-behavior token stream that conditions joint audio-video latent generation. Behavior directives are short language tokens, so they add negligible cost to the token-causal path. The resulting behavior is realized visually (posture, gesture, gaze, interaction with the scene) and acoustically (prosody, non-verbal sound), keeping speech, behavior, and appearance synchronized before decoding rather than aligned afterwards.

# 4 Version Comparison

Table 1 summarizes the changes across versions. The end-to-end streaming formulation, the resolution and frame rate, the latency budget, and the serving topology are inherited from v0.2; v0.3 changes the *conceptual framing* (world + event stream), the *pretraining objective* (predict how a world streams forward), and the *agent's expressive range* (speech plus free-form behavior).

**Table 1** Summary of the Wan-Streamer versions. Emphasized cells indicate what v0.3 changes; unemphasized cells indicate what it preserves from v0.2.

| Aspect | v0.1 | v0.2 | v0.3 |
|---|---|---|---|
| Framing | End-to-end streaming A/V interaction | Same, higher resolution | **Video = world + event stream** |
| Core objective | Interaction data | Interaction data | **General-purpose world-event pretraining, then specialization** |
| Output resolution | 192×336 | 640×368 | 640×368 |
| Frame rate | 25 FPS | 25 FPS | 25 FPS |
| Model-side latency | ∼200 ms | ∼200 ms | ∼200 ms |
| Total latency | ∼550 ms (350 ms network) | ∼550 ms (350 ms network) | ∼550 ms (350 ms network) |
| Serving | Thinker + single-GPU performer | Thinker + Ulysses context-parallel performer | Same as v0.2 |
| Agent event stream | Speech + visible listening behavior | Speech + listening, higher fidelity | **Speech + open-vocabulary free-form behavior** |
| Behavior format | Implicit listening/idle motion | Implicit listening/idle motion | **Role-play with parenthesized behavior directives** |

# 5 Experiments

**Latency and runtime protocol.** We use the same response boundary as v0.1/v0.2. Model-side signal-to-signal latency starts when a 160 ms user streaming unit is available to the thinker and ends when the corresponding audio-video response unit has been decoded for emission. Because the serving topology is inherited unchanged from v0.2, v0.3 keeps approximately 200 ms model-side latency while producing 640×368 video at 25 FPS, and approximately 550 ms total interaction latency with a 350 ms bidirectional network budget. We keep the network term as an external deployment assumption so the comparison isolates the model-side path, and we report the v0.3 runtime at the same boundary used in prior versions [11–15].

**Qualitative behavior observations.** In generated 640×368 conversations, Wan-Streamer v0.3 speaks while performing open-vocabulary behavior in real time. Behavior directives in the language stream are realized as coherent facial expressions, posture changes, responses to sounds, and interactions with nearby objects. These actions remain synchronized with speech, grounded in the configured scene, and consistent with the character's identity and appearance. Between explicit behavior events, the agent maintains stable idle and listening behavior. Wan-Streamer v0.3 therefore expands the agent's expressive range beyond speech and implicit listening while retaining the same low-latency streaming setting as v0.2.

# 6 Conclusion

Wan-Streamer v0.3 reframes native-streaming generation as a **world + event stream** decomposition: persistent world context plus everything that changes over time. This supports general-purpose video pretraining and transfer to roaming, audio-visual interaction, and embodied manipulation. We instantiate this framework through downstream post-training for real-time full-duplex audio-visual interaction. In this instantiation, VLA-like multimodal understanding maps streaming user input to language-form speech and open-vocabulary behavior actions, which condition synchronized audio-video generation. Wan-Streamer v0.3 retains v0.2's modeling contract and serving topology, preserving 640×368 video at 25 FPS, with approximately 200 ms model-side and 550 ms total interaction latency.

# References


[1] Lianghua Huang, Zhi-Fan Wu, Wei Wang, Yupeng Shi, Mengyang Feng, Junjie He, Chen-Wei Xie, Yu Liu, Jingren Zhou, Ang Wang, Bang Zhang, Baole Ai, Chen Liang, Cheng Yu, Chongyang Zhong, Jinwei Qi, Kai Zhu, Pandeng Li, Peng Zhang, Wenyuan Zhang, Xinhua Cheng, Yitong Huang, Yun Zheng, and Zoubin Bi. Wan-Streamer v0.1: End-to-end Real-time Interactive Foundation Models, June 2026. URL https://arxiv.org/abs/2606.25041. Submitted on 23 Jun 2026.

[2] Lianghua Huang, Zhi-Fan Wu, Yupeng Shi, Wei Wang, Mengyang Feng, Junjie He, Chen-Wei Xie, Yu Liu, Jingren Zhou, Ang Wang, Bang Zhang, Baole Ai, Chen Liang, Cheng Yu, Chongyang Zhong, Jinwei Qi, Kai Zhu, Pandeng Li, Peng Zhang, Wenyuan Zhang, Xinhua Cheng, Yitong Huang, Yun Zheng, Yuxiang Bao, Yuzheng Wang, and Zoubin Bi. Wan-Streamer v0.2: Higher Resolution, Same Latency, July 2026. URL https://arxiv.org/abs/2607.04443. Submitted on 5 Jul 2026.

[3] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

[4] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

[5] Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

[6] Xinhua Cheng, Haiyang Zhou, Wangbo Yu, Tanghui Jia, Bin Lin, Yunyang Ge, Weiqi Li, and Li Yuan. 360explorer: Exploring 4d controllable world in panoramic videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3300–3308, 2026.

[7] Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

[8] Tenglong Ao. Body of her: A preliminary study on end-to-end humanoid agent. *arXiv preprint arXiv:2408.02879*, 2024.

[9] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

[10] Lizhen Wang, Yongming Zhu, Zhipeng Ge, Youwei Zheng, Longhao Zhang, Tianshu Hu, Shiyang Qin, et al. Flowact-r1: Towards interactive humanoid video generation. *arXiv preprint arXiv:2601.10103*, 2026.

[11] ByteDance Seed Team. Doubao realtime voice model. https://seed.bytedance.com/en/realtime_voice, 2025. Model page, January 20, 2025.

[12] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Blog post, May 13, 2024.

[13] Hume AI. Introducing evi 3: The world's most realistic and instructible speech-language model. https://www.hume.ai/blog/introducing-evi-3, 2025. Blog post, 2025.

[14] Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. Streamavatar: Streaming diffusion models for real-time interactive human avatars. *arXiv preprint arXiv:2512.22065*, 2025.

[15] Chunyu Li, Jiaye Li, Ruiqiao Mei, Haoyuan Xia, Hao Zhu, Jingdong Wang, and Siyu Zhu. Hallo-live: Real-time streaming joint audio-video avatar generation with asynchronous dual-stream and human-centric preference distillation. *arXiv preprint arXiv:2604.23632*, 2026.

# Appendix

# A Contributions and Acknowledgements

## A.1 Core Contributors

Lianghua Huang*, Zhi-Fan Wu, Yupeng Shi, Wei Wang, Mengyang Feng, Cheng Yu, Chen Liang, Junjie He, Chen-Wei Xie, Yu Liu, and Jingren Zhou.

## A.2 Contributors

Contributors are listed alphabetically by first name: Ang Wang, Bang Zhang, Baole Ai, Chongyang Zhong, Jinwei Qi, Kai Zhu, Pandeng Li, Peng Zhang, Wenyuan Zhang, Xinhua Cheng, Yitong Huang, Yun Zheng, Yuxiang Bao, Yuzheng Wang, Zhiwei Lin, and Zoubin Bi.

*Corresponding author: lianghua.huang.cs@gmail.com.