vllm-omni meeting notes

vLLM-Omni Meeting Notes - Chinese

This meeting is for vllm-project/vllm-omni design reviews and technical discussions:

- Meeting time: 11:30 - 12:30 (UTC+8) Wednesday. (Convert to your timezone)
11:30 - 12:30 (UTC+8) Friday (on demand)
- **Meeting link:** Click here
- **Github:** https://github.com/vllm-project/vllm-omni
- **Website:** https://docs.vllm.ai/projects/vllm-omni/en/latest/
- **Meeting Notes:** https://tinyurl.com/vllm-omni-meeting
- **Release plan:** Click here

Note: This meeting invite **is intended for the Chinese-speaking audiences** to facilitate the timezone. To participate in the meetings intended for English-speaking audiences, please check here.

2026-06-26 Next Meeting

Meeting: vLLM-Omni Meeting 2026-06-26 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)

- Canlin Guo (@gcanlin, MOSS)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Huawei SG, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)
- Miguel Vieira Pereira (@Miguel0312, Huawei France)
- Ruirui Yang(@R2-Y, wechat:luckyyyRs)
- Tianao Cao(@cta-11, Huawei)
- Gaël Glorian (@gglorian, Huawei France, Wechat: Gaël Glorian)
- Wan Tsz Kin(@Numberwan, Huawei, wechat: wantszkin2003)
- Boao Shi(@bowieshi, HKU, wechat: aoibosh)
- Harenome Ranaivoarivony (@harenome, Huawei France, Wechat: Harenome 安仁)
- Weiming Liao(@FayeSpica, Wechat: NineTwelve)
- Hyoseop Song(@loveysuby, Wechat: 宋孝燮)

Agenda:

1. [RFC4378 + PR4529]AURA streaming input
2. [PR4448] Prefetch kv in ar+dit case
3.

Meeting Notes:

2026-06-24 Meeting

Meeting: vLLM-Omni Meeting 2026-06-24 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin, MOSS)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)

- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Huawei SG, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)
- Miguel Vieira Pereira (@Miguel0312, Huawei France)
- Ruirui Yang(@R2-Y, wechat:luckyyyRs)
- Tianao Cao(@cta-11, Huawei)
- Gaël Glorian (@gglorian, Huawei France, Wechat: Gaël Glorian)
- Wan Tsz Kin(@Numberwan, Huawei, wechat: wantszkin2003)
- Boao Shi(@bowieshi, HKU, wechat: aoibosh)
- Harenome Ranaivoarivony (@harenome, Huawei France, Wechat: Harenome 安仁)
- Weiming Liao(@FayeSpica, Wechat: NineTwelve)
- Hyoseop Song(@loveysuby, Wechat: 宋孝燮)

Agenda:

4. RFC4617/PR4618 Harenome
5. PR4433+RFC1601 Boao Shi
6. https://github.com/vllm-project/vllm-omni/pull/3942 paralism adstraction bjf-frz
7. [PR]: Unified VllmOmniConfig :Phase2 Fuyin Wang
8. [RFC]: Unified Data Parallelism for Cosmos3

Meeting notes:

1.

Recording: Click here

2026-06-17 Meeting

Meeting: vLLM-Omni Meeting 2026-06-17 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)

- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)

- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)
- Miguel Vieira Pereira (@Miguel0312, Huawei France)
- Ruirui Yang(@R2-Y, wechat:luckyyyRs)
- Tianao Cao(@cta-11, Huawei)
- Gaël Glorian (@gglorian, Huawei France, Wechat: Gaël Glorian)
- Wan Tsz Kin(@Numberwan, Huawei, wechat: wantszkin2003)
- Boao Shi(@bowieshi, HKU, wechat: aoibosh)
- Harenome Ranaivoarivony (@harenome, Huawei France, Wechat: Harenome 安仁)
- Weiming Liao(@FayeSpica, Wechat: NineTwelve)
- Hyoseop Song(@loveysuby, wechat: 宋孝燮)

Agenda:

1. [RFC]: Unified VllmOmniConfig Fuyin Wang
2. [RFC #4366]: Unified KV Cache Management for the AR-Diffusion Engine (BDE)
3. [RFC3715 + PR4493] Plugin-based sparse attention interface + per-forward DiT @harenome)

Meeting notes:

2026-06-12 Meeting

Meeting: vLLM-Omni Meeting 2026-06-12 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)

- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)
- Miguel Vieira Pereira (@Miguel0312, Huawei France)
- Ruirui Yang(@R2-Y, wechat:luckyyyRs)
- Tianao Cao(@cta-11, Huawei)
- Gaël Glorian (@gglorian, Huawei France, Wechat: Gaël Glorian)
- Wan Tsz Kin(@Numberwan, Huawei, wechat: wantszkin2003)
- Boao Shi(@bowieshi, HKU, wechat: aoibosh)

Agenda:

1. [RFC3865] Omni-Replica + vLLM-DP + EP/EPLB
2. [PR4079] [Refactor] Migrate diffusion prompt-list batching to request-level scheduler batching
3. [RFC4366] AR-DiT KV Management Bob Zhou
4. [PR4041] [Feature] Add HunyuanImage3 DiT grouped step batching

Meeting notes:

Recording: Click here

2026-06-10 Meeting

Meeting: vLLM-Omni Meeting 2026-06-10 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)

- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)
- Miguel Vieira Pereira (@Miguel0312, Huawei France)
- Ruirui Yang(@R2-Y, wechat:luckyyyRs)
- Tianao Cao(@cta-11, Huawei)
- Gaël Glorian (@gglorian, Huawei France, Wechat: Gaël Glorian)
- Wan Tsz Kin(@Numberwan, Huawei, wechat: wantszkin2003)

Agenda:

5. [PR4192] [Misc] LVSA showcase (training-free block-sparse attention) (Contributor from Europe, reorder due to hard time zone)
6. [PR4225] Base class for dit pipelines with unified parameter declaration[1/N]
7. [RFC3978] Inter-Request DiT Cache with Semantic Reuse
8. [RFC4246] New model support: AURA
9. [PR4302] Add support for PipeFusion and integrate it into Wan 2.2

Recording: Click here

2026-06-03 Meeting

Meeting: vLLM-Omni Meeting 2026-06-03 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)

- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)
- Miguel Vieira Pereira (@Miguel0312, Huawei France)
- Wan Tsz Kin(@Numberwan, Huawei, wechat: wantszkin2003)

Agenda:

10. [PR3855] Refactor Omni Stage Runtime and Distributed Replica Control Plane
11. [PR3701] Implement the Lingbot World Fast Pipeline to work in vllm-omni @Miguel0312
12. New maintainer nomination: alex & minghui

13. [RFC 3935] Skill RFC
14. [PR 3208] support qwen-image disaggregated encoder

Meeting Notes:

2026-05-27 Meeting

Meeting: vLLM-Omni Meeting 2026-05-27 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)

- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)
- Wan Tsz Kin(@Numberwan, Huawei, wechat: wantszkin2003)

Agenda:

1. [PR#3860] Add profiling-based diffusion batching and keyed DRR scheduling
2. [RFC#1987]: World Model Support
3. [PR3710] Lance(Bytedance)
4. [PR 2126] AMD Micro world model, model preparation approach Pull requests · vllm-project/vllm-omni
    a. Just one question: It depends on two model repositories, what is a preferred way to support downloading of two models?
5. [PR 3628] Benchmark data statistics for each stage of omni models

Meeting Notes:

Recording: Click here

2026-05-20 Meeting

Meeting: vLLM-Omni Meeting 2026-05-20 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)

- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Zhang Zhen (@zzhang-fr, Huawei France, Wechat: zen)

Agenda:

1. Chunked prefill code review support chunked prefill for qwen3-omni by LJH-LBJ · Pull Request #2997 · vllm-project/vllm-omni - Liu Junhong
2. LVSA RFC3115 introduction and integration discussion (with RFC3715) - Zhang Zhen

3. [RFC #3735] Model-Aware Argument Default Resolution - WU Hang
4. [RFC#3545] Extend Prometheus

Meeting notes:

Recording: Click here

2026-05-15 Meeting

Meeting: vLLM-Omni Meeting 2026-05-15 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)

- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu(AntGroup)
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)
- Meng Chen(Wechat: Dameccyyy)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)

Agenda:

1. [RFC#3554] Robotics Integrations
2. [PR#3614] Streaming Input for Async Chunk
3. [PR#3569] Integrate OmniCoordinator into stage engine pipeline
4. RFC#3632 Streaming diffusion video generation output & mid-way prompt update
5. [RFC#3545] Extend Prometheus

Meeting Notes:

Recording: Click here

2026-05-13 Meeting

Meeting: vLLM-Omni Meeting 2026-05-13 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)

- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@nussejzz, Tencent, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu()
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)
- Zejian Wang(@knitcapcat, AMD, wechat: 15222838158)
- Bincheng Kang(@sphinxkkkbc, wechat: sphinxkkkbc)
- Xiaohui Mu(@Flink-ddd, Wechat: iZzy07Zoey12)

Agenda:

1. v0.22.0 release plan
2. [RFC#3535] Qwen3-TTS high-throughput hardening
3. [FEATURE#3562]hunyuan image test coverage
4. [RFC #3550] [Refactor]: Unify diffusion request identity around request_id
5. Review of bug severity classification
6. [PR#1742] Mori-IO Transfer Connector

Meeting Notes:

Recording: Click here

2026-05-06 Meeting

Meeting: vLLM-Omni Meeting 2026-05-06 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)

- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu()
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)

Agenda:

- Monthly committer nominations: yuanheng zhao & ruixiang ma
- Version release key issues discussion

Meeting Notes:

Recording: Click here

2026-04-29 Meeting

Meeting: vLLM-Omni Meeting 2026-04-29 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@lsotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)

- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu()
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)
- Su Zhengyuan (@timzsu, NUS)
- Chen Cheng (@ischencheng, UPenn)

Agenda:

- [RFC #3131] let TSS/omni audio encoder use `vllm.vllm_flash_attn`
- [RFC #3186] Kernel Optimization for Diffusion DiT and MoE LLM
- [PR #2751] Wan2.2-S2V model enabling
- [RFC #3163] TTFB scaling under concurrency
- [PR #3157] Realtime video input/output(Wan2.2 Causal)
- [RFC #3228] Prometheus Metrics Support

Meeting notes:

2026-04-22 Meeting

Meeting: vLLM-Omni Meeting 2026-04-22 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)

- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu()
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)

Agenda:

1.[PR 2396] [FEAT] support multi-stage deployment

2. Vllm-omni 0.20.0 release update

Meeting notes:

2026-04-15 Meeting

Meeting: vLLM-Omni Meeting 2026-04-15 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)

- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu()
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)
- Rustam Khadipash (@hadipash, Huawei)

Agenda:

1. [RFC #2363][PR-2322] Add Pipeline Parallelism into vLLM-Omni
2. [RFC #2403][PR-2724] Diffusers Backend Integration for Extended Model Coverage
3. [RFC #2366] L5 Reliability Test

Meeting notes:

2 Diffusers Backend Integration for Extended Model Coverage

- The current design is different from how vllm integrates HF transformers
- Discuss with the diffusers team about how they would like to be integrated
- Check out the following relevant designs
    - Vllm + transformers
    - Sglang + diffusers

3 L5 reliability test

Add four scenarios: abnormal input; OOM; process kill; runtime teardown (container kill)

2026-04-08 Meeting

Meeting: vLLM-Omni Meeting 2026-04-08 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)

- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: ray_u30)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)
- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng(AntGroup, Wechat: Johney_Zheng)
- Lianhao Xu()
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)

Agenda:

4. [RFC #2379][PR-2507][WIP]vllm-omni CUDA IPC support
5. [PR-2396]vllm-omni Stages scale out(support multi-stage deployment)
6. New HW support (Moore Threads GPU), tracked by [issue #2347]
7. FA3 FP8 Li Shunyang顺阳
8. [Test] CI Operation Rules and Rectification Plan

Meeting notes:

2026-04-01 Meeting

Meeting: vLLM-Omni Meeting 2026-04-01 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)
- Michael Qiu(@Michael Qiu, AntGroup, Wechat: SEU_08005325)

- Ngai Fai Ng (@chickeyton, Huawei, Wechat: chicketyon)
- Wengang Zheng()
- Lianhao Xu()
- Zhang Jian (@zhangj1an, Wechat: lavendromeda)

Agenda:

9. [PR #2006] [WIP]Refactor StageDiffusionClient and StageEngineCoreClient
10. [PR #1822, PR #2301] Add Ming-flash-omni-2.0 Thinker
11. [PR #2208] Add session based audio streaming input
12. vllm-omni Q2 roadmap discussion
13. vllm-omni new maintainer nomination

Meeting Notes:

2026-03-25 Meeting

Meeting: vLLM-Omni Meeting 2026-03-25 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)

- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)

Agenda:

1. [PR #2020] [Entrypoint][Refactor]Stage CLI Refactor
2. [PR #1822] Model Support Ming-flash-omni-2.0
3. [PR #2134] [Model] Adapt Wan2.2-I2V-A14B via LightX2V offline conversion path

Meeting Notes:

Recording: Click here

2026-03-18 Meeting

Meeting: vLLM-Omni Meeting 2026-03-18 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)

- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)
- Wenkang Xu(@Neherio, Huawei, Wechat: xwk980327)

Agenda:

1. [RFC 1777] Add offline quantized W4A16 model support through Intel AutoRound
2. [RFC 1949] Dynamic Expert Parallel LoadBalance for DiT-MoE
3. [PR1908] [Entrypoint][Refactor] vLLM-Omni Entrypoint Refactoring

Meeting Notes:

3. [PR1908]: finished, ready to merge today. Write a plan for follow-up todos.

Recording: Click here

2026-03-13 Meeting

Meeting: vLLM-Omni Meeting 2026-03-13 11:30 AM (UTC+8)
Join: Click here

Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@lsotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)

- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)

Agenda:

1. [RFC 1546] [PR 1555] refractor communication layer
2. [PR 1769] Support step-level request abort for diffusion models
3. [PR 1826, RFC 1860] default stage config dispatch based on workload

Meeting Notes:

Recording: Click here

2026-03-11 Meeting

Meeting: vLLM-Omni Meeting 2026-03-11 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)

- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
- Lin Yueqian (@linyueqian, Duke)
- Jiaping Wu（@ElleElleWu, Huawei,Wechat:Woo-GNIPAIJ)

Agenda:

1. [PR 1689]: [Model] Extend NPU support for HunyuanImage3 Diffusion Model
2. [PR 1559, 1558, 1560, 1561] Omni stage crash/proc exit capture
4. [PR 1526] update GpuMemoryMonitor to DeviceMemoryMonitor for all HW
5. [RFC 1795] TTS Development Roadmap - March 2026
6. [PR 1721] [CI] init intel ci dispatch in buildkite
7. [RFC 1763] Unified Quantization framework

Meeting notes:

Recording: Click here

2026-03-06 Meeting

Meeting: vLLM-Omni Meeting 2026-03-06 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
- Chendi Xue(@xuechendi, Intel)
-

Agenda:

8. [RFC 1568]:[RFC]: Discussing the extension of attention backend
9. [PR #1328]: High-Performance MoT (Mixture-of-Tokens) Kernels

Meeting Notes:

Recording: Click here

2026-03-04 Meeting

Meeting: vLLM-Omni Meeting 2026-03-04 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)
- Bo Dong (@dongbo910220, wechat: billbalack)
-

Agenda:

1. [PR1115]: Config Refactor[1/N] Model pipeline Configuration System.

2. [RFC, design doc]:Add Ulysses advanced_uaa mode
3. [RFC 1595]: Decouple Model-Specific Logic from Model Runner
4. [RFC 1601]: Decouple Multimodal Output Channel & Simplify Output Processor
5. Nomination of new committers for hardware plugin system, diffusion and tts

Meeting notes:

1. [PR1115]: Update PR information and future plan in a clear manner. Move pipeline_yaml into the respective model folder to facilitate integration.
2. [RFC]: Proceed with the PR according to the documentation.
3. [RFC 1595]: Migrate Model Runner v2 firsts, base model as Qwen 3 omni
4. [RFC 1601]: Need further discuss with vllm features, i.e. extra output and see if can collaborate
5. Nomination of new committers for hardware plugin system, diffusion and tts are @gcanlin**,** @wtomin, and @linyueqian

Recording: Click here

2026-02-25 Meeting

Meeting: vLLM-Omni Meeting 2026-02-25 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)

- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)
- Yuanheng Zhao (@yuanheng-zhao)
- Yiqi Xue (@Vivo50E, wechat: Xueey7)
- Yi Liu(yi4.liu@intel.com, Intel, wechat: Randall_ku)
- Shanshan Shen (@shen-shanshan, Huawei Open Source)

Agenda:

1. **[RFC] Implementation of Deterministic Sleep Mode and ACK Protocol**
2. Develop vLLM-omni Docker image to accelerate CI
3. [RFC]: Rebase Additional Information into Model Intermediate Buffer
4. [RFC]: vLLM-Omni Multi-Stage CFG Support
5. [RFC]: Support KV Cache CPU Offloading #1150

Meeting Notes:

2. rc version will release docker image cut on the rebase commit, which can be used for future CIs
3. Additional information rebase may affect async chunk and the interface need further discussion
4. Need wrap the cfg related code to some function, minimize the change in Omni

Recording: Click here

2026-02-11 Meeting

Meeting: vLLM-Omni Meeting 2026-02-11 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)

- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)

Agenda:

1. [RFC] Continuous Diffusion Model Acceleration Support
2. Nomination for new maintainer@vllm-omni
3. [PR + RFC] Add Omni Model Performance Benchmark Test
4. [RFC]: Refactor engine/runner/pipeline to support step-wise and continuos batching

Meeting Notes:

1. [RFC] Continuous Diffusion Model Acceleration Support
   Refine the guidance for feature x feature validation.
2. [PR + RFC] Add Omni Model Performance Benchmark Test
   1) Support for diffusion-related models will be added in the next version of VLLM Bench, with plans to include diffusion extensions.
   2) Add a performance data validation module (benchmark comparison).
3. Nomination for new maintainer@vllm-omni:
   After discussion, we confirm Zhipeng Wang @princepride , as the new committer.

4. Provide an initial version of the pipeline PR.The step-level interfaces should be introduced in a backward-compatible way, without affecting fast Day-0 model onboarding and support.

Recording: Click here

2026-02-06 Meeting

Meeting: vLLM-Omni Meeting 2026-02-06 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)

- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)

Agenda:

1. [https://github.com/vllm-project/vllm-omni/issues/1154] Support circular AR+DIT
2. [RFC 1184] Support Prefix Caching for AR omni part
3. [RFC 1030] Custom Op for diffusion
4. [RFC 1218] L2 & L3 Test Case Stratification Design for Omni Model

Meeting notes:

1. [RFC 1030] Custom Op for diffusion: 1.for triton kernel, perhaps we wrap it for different platforms, e.g. qwen_joint_rope_amd. 2. Diffusion torch compile integration: not in the near term. 3. For kernels in vllm but don't have aligned interface, we will try our own triton kernel first.
2. [RFC 1218] L2 & L3 Test Case Stratification Design for Omni Model: Explain the use case layering logic and script implementation design at L2 and L3 levels. A PR implementation is planned by next weekend, including corresponding use case templates and documentation

Recording: click here

2026-02-04 Meeting

Meeting: vLLM-Omni Meeting 2026-02-04 11:30 AM (UTC+8)
Join: Click here

Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)
- Xinyu Chen (@xinyu-intel, Intel, wechat: jitmatrix)
- Ding Zuhao (@NUS, WeChat: jzz13855888819)

Agenda:

1. [RFC 1057]: Q1 Diffusion Quantization Support
2. [RFC 400 & PR 1167]: CI Five-level design @yenuo Cong Wang
3. [RFC 1146]: Support Fused MoT Kernels for Bagel
4. Release plan for Feb 2026 (@Gaohan123) **Click for details**
5. [PR 1046] Set up platform dependent package installation

Meeting Notes:

1. [RFC 1057]:

a. **Hardware:** NPU is still just in research for now. AMD needs more sweat equity to get vLLM running right.
b. **Compression:** Need to huddle with stakeholders before pulling the trigger on the LLM compressor. AWQ is looking like a no-go for Diffusion (needs testing), and GGUF is going to be more troublesome since we need to coordinate with three different parties to get it working with transformers. No plan for Int4 support.

c. **APIs:** We need to be careful with the interface design now—it's going to get messy later if we don't make it user-friendly from the start.
d. **Project structure:** Put quantization in the diffusion folder for now. If a specific model needs its own custom ops, just keep that code inside that model's folder to keep things clean.
e. **Next steps:** More testing and digging into mainstream benchmark methods to see how we should be measuring.

2. PR 1167: CI Five-level design. The overall framework is okay. We need to provide a template for the L2 & L3 E2E test.

3. RFC 1146: Following vLLM's MoE module design, abstract base classes and quantization config classes will be introduced under `diffusion/layers/mot/`, together with Triton kernels. The actual quantization logic will be implemented later under `diffusion/layers/quantization/`.

4. Release plan for Feb 2026. OK to follow the plan. Already updated in the head of file.

Recording: Click here

2026-01-28 Meeting

Meeting: vLLM-Omni Meeting 2026-01-28 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)

- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)

Agenda:

1. [RFC 967]: [Draft][RFC]: vLLM-Omni Entrypoint Refactoring
2. [PR #774]: [Hardware] Support platforms and plugin system
3. [RFC#874] Refactor engine/runner/pipeline to support step-wise and continuous batching
4. [RFC #996] Version Align with vLLM
5. [RFC #750] [MODEL] Mimo-Audio support, short report

Meeting Notes:

1. [RFC #750] Wrap the special MIMO-audio code into a function and protect it with unit tests.
2. [RFC#874] Start by refactoring the diffusion pipeline, then integrate the same-resolution continuous batching PR.
3. [RFC 967] The main discussion focused on the differences in process management compared to existing designs. The next step will be to further break down tasks and proceed with the refactoring.

Recording: Clink here

2026-01-21 Meeting

Meeting: vLLM-Omni Meeting 2026-01-21 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)

- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)

Agenda:

1. [RFC #353 & PR #716]: YuanrongConnector — OmniConnector Implementation for vLLM-Omni
2. [RFC #850]: CFG Parallelism Abstraction.
3. [RFC https://github.com/vllm-project/vllm-omni/issues/870] Support DP Stage
4. [PR #837]: optional modality aware scheduler and 2 benchmark tests.
5. [PR #847] Rebase to 0.14.0

Meeting Notes:

2026-01-14 Meeting

Meeting: vLLM-Omni Meeting 2026-01-14 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, vLLM-Omni, wechat: XU-XMS)

- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)
- Canlin Guo (@gcanlin)
- Huang Zeyu (@fhfuih, Huawei, wechat:)
- Yang Songlin (@yangsonglin13, Huawei, wechat: Sunny2012Yang03)
- Shijin Zhang(@Xiaomi, wechat:RsyncDoven_Imp)
- Jingan Zhou (@AndyZhou952)

Agenda:

1. [RFC #353 & PR #716]: YuanrongConnector — OmniConnector Implementation for vLLM-Omni
2. [RFC 427 & 701, design doc]: Interface refactor to supported batch diffusion request
3. [PR #727]: Support async computation and communication across stages by chunks
4. [PR #774]: [Hardware] Support platforms and plugin system
5. [PR #719]: [CI] Pytest markers **&&** [Design doc] CI Execution plan part Cong Wang
6. [PR#750]: [MODEL] XiaomiMiMo/MiMo-Audio-7B-Instruct support
7. [PR #758] Diffusion LoRA Adapter Support (PEFT compatible) for vLLM alignment
8. [RFC #771] Decoupling Text Encoder in Diffusion Model Inference Pipeline
9. [PR #709]: Diffusion torch profiler

Meeting notes:

1. [RFC #353 & PR #716]: Refactor the Connector into a general-purpose component to enable GPU deployment, and supplement the performance test data in the GPU environment in the PR.
2. [RFC 427 & 701, design doc]: ongoing impl at https://github.com/fhfuih/vllm-omni/pull/1/. Next step is to ensure all models works well, and add some Unit Test.
4. [PR #774]: This PR requires attention from all hardware vendors and collaborative efforts to complete the corresponding platform support.
6. [PR#750]: Move the function **_batch_mm_kwargs_from_scheduler** out of **gpu_model_runner.py** and place it elsewhere, or keep it as is (with sufficiently clear justification provided). Next steps: batching strategy, CUDA Graph implementation, and performance metrics.
5. CI: New marks are good to merge; CI execution plan: Need to decide the marks for features/functions later; Buildkite daily CI setting needs rogerw@vllm.ai.
7. [PR #758] Diffusion LoRA Adapter Support (PEFT compatible) for vLLM alignment: current addition/design looks good on the whole, have asked vllm-lora PoC to follow up. Next steps: finish testing, doc, tidying up weight loading etc.
8. [RFC #771] we all agree that we need to separate text encoder, but we need more discuss on details like which engine to use

Recording: Click here

2026-01-07 Meeting

Meeting: vLLM-Omni Meeting 2026-01-07 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, Huawei, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Zhu Yufeng (THU,GT, wechat: windy-zhuyufeng)

Agenda:

1. vLLM-Omni CI design (markers, CI tests design, full tests design) CI design issue#400 Test marks PR#577 vLLM-Omni CI/CD design Cong Wang
2. vLLM-Omni Q1 Roadmap release
3. [PR669] Support Qwen3 Omni talker cudagraph

Meeting notes:

1. Pytest markers:
    - Using multi-gpu decorator instead of multi gpu marker
    - Consider the nightly full test suite
    - Remember to update the English doc / PR intro.

   CI tests:
    - Store the full test suite in the same folder as the current CI test, but diff python files.
2. [PR669] Refactor talker mtp cudagraph for better generalization in another PR. Support talker bs>1 ASAP.

Recording: Click here

2025-12-31 Meeting (Happy New Year!)

Meeting: vLLM-Omni Meeting 2025-12-31 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, Huawei, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechatd: Zjt_Tete)
- Ziming Huang(@Alibaba Cloud, wechat:hzim2000)
- Li Shunyang(@NUS, wechat: LSY2717785144)

Agenda:

1. vLLM-Omni December progress review and 0.12.0 release discussion
2. [PR486] Support Omni serving abort request
3. [RFC409] Qwen3-Omni deployment
4. [PR551] Make preprocess and forward in the same function
5. [PR367] Basic version of supporting streaming output
6. [Issue481] Support torch profiler
7. [PR319]BagelI&HunyuanImage support

Meeting notes:

1. Agree to release vllm-omni 0.12.0rc1
2. [PR486] Support Omni serving abort request. Check why ci failed, merge after ci fixed.
3. [PR367] Basic version of supporting streaming output. Resolve conflicts and merge ASAP.

Recording: Click here

2025-12-24 Meeting

Meeting: vLLM-Omni Meeting 2025-12-24 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , vLLM-Omni, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, Huawei, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechat: Zjt_Tete)

Agenda:

8. [RFC #308] Support ASDSV for Wan2.1 Video
9. [PR #273] Diffusion Ring Attention support
10. [RFC #442] Log system refactoring
11. [PR 391] [Core] Supports stage abstraction in the diffusion model

Meeting notes:

2025-12-17 Meeting

Meeting: vLLM-Omni Meeting 2025-12-17 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)

- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, Huawei, wechat: XU-XMS)
- Zhipeng Wang(@princepride, wechat: wzp_princepride)
- Jintao Zhang(@THU, UCB, wechat: Zjt_Tete)

Agenda:

1. [RFC#290] Diffusion Chunked Scheduling RFC (Draft)
2. [PR298] Support output modalities control per request
3. [PR189] Diffusion Ulysses-Sequence-Parallelism support
4. [PR340]Adding profiling hooks for omni&vllm&diffusion pipeline
5. TurboDiffusion integration?

Meeting Notes:

1. [RFC#290] Diffusion Chunked Scheduling RFC (Draft)
    a. re-evaluate scheduling granularity—should it be simple "Steps" or defined "Chunks"
    b. Analyze step cost variations across resolutions to determine the scheduling policy for mixed-resolution workloads.

2. [PR298] Support output modalities control per request. It can be merged after resolving the rebase bug.
3. [PR189] : revise the test pipeline to include e2e ulysses sp test, remove ring_degree argument, and refine the documents.
4. [PR340]Adding profiling hooks for omni&vllm&diffusion pipeline
    a.Add documentation explaining how to use logs.
5. TurboDiffusion integration: First add a vllm-omni wrapper to call turbodiffusion.

Recording: Click here

2025-12-10 Meeting

Meeting: vLLM-Omni Meeting 2025-12-10 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)

- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Tun Jian Tan (@tjtanaa, vLLM, wechat:)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)
- Mingshi Xu (@mxuax, Huawei, wechat: XU-XMS)

Agenda:

6. [PR250] Add cache-dit Jason Huang
7. Test design&dev: dir structure, coding style, and current problems. Cong Wang
8. [PR212]Benchmark for Qwen 3 Omni Bob Zhou
9. [PR215]Omni Connector + ray supported wzliu
10. [PR179]TeaCache for Qwen Image.
11. [RFC#252]Context Parallelism (RingAttention) and Parallelism Terminology Alignment
12. [PR259] [WIP]Support Online Serving for Diffusion-only Models

Meeting Notes:

1. Installing Cache-dit with github commit can lead to network problem in mainland, will update to pip package after cache-dit releases a wheel package with `refresh_context` interface support.
2. Test design&dev: UT priority - low; Only test qwen omni2.5 & qwen image for online testing; DIR structure change is fine; Cong Wang output a template for model offline testing.

6. When implementing parallel acceleration for DIT, since it doesn't require using many VLLM interfaces, the conventional understanding of SP from the DIT community is retained. If the difference between SP and CP needs to be clarified later, it should be discussed with the VLLM community. Ring attention will be added after implementing Ulysses. Mingshi XU

7. The current version of online serving for diffusion-only models can be merged first, and the version that integrates with the OmniStage can be continuously integrated in the future. Chenguang Zheng

Recording: Click here

2025-12-03 Meeting

Meeting: vLLM-Omni Meeting 2025-12-03 11:30 AM (UTC+8)
Join: Click here

Attendees:

- Gao Han (@Gaohan123, vLLM-Omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-Omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Cyrus Leung (@DarkLight1337, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)
- Yikun Jiang (@Yikun, vLLM Ascend, wechat: yikunkero)
- Didan Deng (@wtomin, Huawei, wechat: miao958548249)

Agenda:

1. 1230 Roadmap discussion
    a. Core
    b. CI/Tests
    c. Model Supports
    d. Refactor
    e. Docs
    f. Benchmark
    g. Aligned with v0.12.x
2. DiT enhancement
    a. Diffusion Core
    b. Model supports
    c. Acceleration plugin
3. CI tests and Bugfix allocation
4. [RFC]: Diffusion Acceleration API design
5. [PR148] [Core refactor] customize preprocess, remove "request_ids" in model forward
6. [RFC]: Automatic GPU Mem Utilization Tuning

Meeting Notes:

1. [RFC]: Automatic GPU Mem Utilization Tuning:
    a. First, assign memory upper bounds for each stage. Then, set the GPU utilisation based on these.
    b. The approximation should not automatically change the GPU utilisation, but should simply provide suggestions.
2. [RFC]: Diffusion Acceleration API design
    a. Let open-source community to work on cache acceleration. For now, TeaCache already has contributors. @LawJarp-A
    b. Parallel attention interface design: needs to be compatible with attention backends (PR #115).

3. [PR148] [Core refactor] customize preprocess, remove "request_ids" in model forward
    a. Finish Qwen3-omni. Test and wait for merge. Next, work on new interface for model file and yaml file.
4. 1230 Roadmap discussion
   High priority features:
   a. UT/ST of current examples
   b. Online serving for visual generation
   c. Output modality control
   d. Docs Refinement
   e. v0.12.0 vllm refractor
5. Urgent CI tests allocation:
   Qwen2.5-omni online serving: Zhou Taichang
   Qwen3-omni: Yang Ruirui
   Qwen3-omni Online serving: Gao Han
   Qwen-Image: Zhu Jiangyun
   Z-image: Zhu jiangyun/Huang Yongxiang
   Gradio demo: Huang Yongxiang
   Qwen 2.5 offline & NPU support: Guo Canlin
6. Bugfix and CI maintenance allocation:

Week 1	Week 2	Week 3	Week 4
Congw729	yinpeiqi	tzhouam	SamitHuang
David66666	R2-Y	hsliuustc0106	Gaohan123

Week 5	Week 6
gcanlin	natureofnature
ZJY0516	qibaoyuan

Recording: Click here

2025-12-01 Meeting

Meeting: vLLM-Omni Meeting 2025-12-01 11:30 AM (UTC+8)

Join: Click here

Attendees:

- Gao Han (@Gaohan123, vLLM-omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)

Agenda:

1. Discussion for first release and arrangement for further meeting.
2. Blog post PR https://github.com/vllm-project/vllm-project.github.io/pull/123
    a. https://vllm-project-github-6s2gdextt-simon-mos-projects.vercel.app/2025/11/30/vllm-omni.html
3. Release checklist vLLM-Omni Release Checklist
4. [PR115] Diffusion Attention init

Meeting Notes:

1. For pytest markers:
    a. Using multi-gpu decorator instead of multi_gpu_x markers;
    b. Consider the nightly omnidirectional tests
    c. Remember to update the English document / PR intro

    For CI test cases:
    a. Store the omnidirectional tests (全量测试) in the same folder with current CI tests but diff python files.
    b.

Recording: Click here

2025-11-28 Meeting

Meeting: vLLM-Omni Meeting 2025-11-28 11:30 AM (UTC+8)

Join: Click here

Attendees:

- Gao Han (@Gaohan123, vLLM-omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)

- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)

Agenda:
[PR86] Doc system merged and further things
[PR82] Qwen3 Image is ready
[PR89] Add Ascend NPU Backend support for omni model
[PR92] Allow customization process between stages, removing the need for "request_ids" in the model forward.
[PR93]Buildkite + Unit test

Meeting Notes:
1. [PR86] Welcome every contributor to check and modify current doc files if necessary. For wechat groups operation, Roger Wang will help to ask kaichao for reference from vLLM main repo. Further optimization for docs can continuously make progress after first release.
2. After discussion, the first release will be scheduled on Feb 1, 2025 (Mon)
3. [PR92] qwen2.5 omni is done, try to simplify the abstraction of qwen3 omni.
4. [PR89] Qwen2.5-Omni can now run end-to-end on Ascend NPU and generate intelligible human speech. Need to clarify the change of operator impact on GPU. In addition, and will create a dedicated `npu/` directory and temporarily keep all NPU-related code in the vllm-omni repository, with the option of migrating it into a plugin in the future.

Recording: Click here

2025-11-26 Meeting

Meeting: vLLM omni Meeting 2025-11-26 11:30 AM (UTC+8)
Join: Click here
Attendees:
- Gao Han (@Gaohan123, vLLM-omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)

Agenda:
[Issue 74] Simplify model stage config for end users.
[PR55] Qwen3 Omni Support

[PR86] Optimize and supplement docs system
[PR82] Qwen3 Image is ready for review
Meeting Notes:

- Qwen3-omni: Add hidden states output interface in upstream vLLM to support both text/hidden_states output in the future.
- [PR82] unify offline entrypoint: Omni(model_name)
- [PR86] Optimize and supplement docs system. Continue to optimize the content following reviews.
- Simplify model stage: work for a demo first.

Recording: Click here

2025-11-24 Meeting

Meeting: vLLM omni Meeting 2025-11-24 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)

Agenda:
[PR] Qwen-Image init support
[PR75] Script for building whl Cong Wang
[PR55] Qwen3 Omni Support
[PR79] Omni connector
Meeting notes:

- Qwen3 Omni precision issue
  - FusedMoE kernel has some precision issue in both thinker and talker
  - Swap to torch native to see where the biggest error comes from
  - Follow up with Qwen & vLLM team
- [PR75] Building whl
  - Merge first, and maybe resolve the entrypoints overwriting problem later.
  - Add files for vllm_omni.__version__
- [PR79] Omni connector
  - Discussed the potential performance gain through fully disaggregated architecture

        - Small models lacking performance from conserve team's experiences
        - Roger suggests large models are the trends and supports the necessity of disaggregation.
    - [Follow up] Share the slides by attaching with the PR

Recording: Click here

2025-11-21 Meeting

Meeting: vLLM omni Meeting 2025-11-21 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)

Agenda:

- [Feature] Support multimodal inputs with multiple requests PR76
- [Discuss] How to support Qwen-Image
- [CI] Add buildkite on vllm-omni so that it can share the same resource pool with vllm-project/vllm
- [whl] Script for building whl PR75 Cong Wang
- [Readme] https://github.com/vllm-project/vllm-omni/pull/69
-

Meeting notes:

- audio_in_video support for Qwen-omni online
- Create a new folder for a new diffusion model runner under vllm-omni for qwen-image
    - Merge new diffusion model runner and vllm-omni/worker/diffusion model runner if possible
- Roger Wang to onboard buildkite on vllm-omni
    - Simple test to ensure functionality, correctness may still need human involvement.
- [whl] Script for building whl PR75 - using uv setup env, after testing, it can be merged. Write the related doc later; Check CI why it is not auto-run.
- [Feature] Support multimodal inputs with multiple requests PR76 . The overall implementation is ok. Waiting for Roger Wang to review.

Recording: Click here

2025-11-19 Meeting

Meeting: vLLM omni Meeting 2025-11-19 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Jeff Ma (@majunze2001, WeChat: m2261257350)

Agenda:

- Discuss – Plan for wheel release; Add bot for project (for PR automerge, wheel auto-releasement); Reviewer auto-assign Cong Wang
- RFC: simplify model stage config for end users
- Discuss - How do I debug one stage in a multi-stage pipeline, especially an intermediate stage?

Meeting notes:

1. Wheel release:
    - Do not pre-install dependencies for examples/*. Let users install them based on the README file.
    - Change `pyproject.toml` to include the vllm dependencies (don't include vllm)
    - GitHub Actions resources for a free account only have 500 MB of storage. rogerw@vllm.ai will contact Kaichao.

    Adding bot:

    - Two use cases: automerge and auto-building wheel.
        - For the automerge task, Roger recommends that the reviewer click the "auto-merge" button after reviewing instead of using the bot to automerge.
        - For the auto-building wheel task, Roger recommends that we provide the script to build the wheel and a stable wheel file after the new version is released.

    Codeowners:

    - Now the CODEOWNERS file is not affected, and we need the repository owner to enable it. Will talk with @gaohan later to enable this feature.
    - Assign corresponding reviewers for different folders.

2. RFC: simplify model stage config for end users, is a right direction. It can continue to work on. But it is not necessary for the first release.

Recording: Click here

2025-11-17 Meeting

Meeting: vLLM omni Meeting 2025-11-17 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, vLLM-omni, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, vLLM-omni, wechat: hsliuustc)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@Isotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)
- Yang Ruiui (@R2-Y, Huawei)

Agenda:

- [Feature] Support online inference PR64
- [Feature] add support for Qwen3-omni PR55
- Preparing for 0.11.0rc1, expected to be released before the end of this month, this week could be better
    - Docs are in a mess
        - 1) requiring reorganization
        - 2) main arch docs are missing or out-of-date
    - Scripts folder to be deleted
    - Docker installation support
    - No benchmarks: vllm bench serve before 0.11.0 release
    - Readme needs to be discussed: **following from vLLM main project**
        - title for vLLM-omni
            - Option 1(current): Multi-modal Extension for vLLM
            - Option 2: Easy, fast, and cheap Omni-modality models serving for everyone
        - Project introduction:
            - Option 1(current): vLLM-omni supports multi-modality models inference and serving with non-autoregressive structures and non-textual outputs, extending vLLM beyond traditional text-based, autoregressive generation.
            - Option 2: A high-throughput and memory efficient inference and serving engine for Omni-modality models
        - About:

        - vLLM-omni is fast with
            a. Seamlessly AR support with vLLM
            b. Diffusion model execution acceleration by xxx
            c. Pipeline stage execution overlapping
            d. Fully disaggregation based on omniConnector and dynamic resource allocation
        - vLLM-omni is flexible and easy to use with
        - vLLM-omni supports seamlessly supports most popular open-source models on HuggingFace, including:
            a. Qwen
            b. Hunyuan
            c. SD
            d. flux
    - Getting Started: following from vLLM main project
    - Contributing: following from vLLM main project
    - vLLM-omni logo: shall we use qwen-image to help us generate the logo?
- Meetup slides preparation for vLLM-omni repo

Meeting notes:

1. [Feature] Support online inference PR64, the implementation is approved. Ready to merge.
2. vllm-omni plans to take the consistent version number with the main vLLM it depends on. For example, now it depends on vLLM of v0.11.0. So the version of vllm-omni is v0.11.0rc1 for primary release. And v0.11.0 for stable release.
3. Based on CI workflow, we can support installation of pip wheel.
4. [Feature] add support for Qwen3-omni PR55, GenerationWorker, GenerationModelRunner can be base classes inherited by DiffusionWorker, DiffusionModelRunner. Other implementations make sense.
5. For issue58 and issue65, @ZJY0516 and @lsotr0py will instantiate the interface of DiT with support of Qwen-image.
6. For preparation of first release 0.11.0rc1, follow TODOs in the agenda above.

Recording: Click here

2025-11-14 Meeting

Meeting: vLLM omni Meeting 2025-11-14 11:30 AM (UTC+8)
Join: Click here
Attendees:

- Gao Han (@Gaohan123, Huawei, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, Huawei, wechat:)

- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@lsotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)
- Baoyuan Qi(@qibaoyuan, XiaoMi, WeChat:deepthink2055)

Agenda:

- [fix] adapt hidden state device(**https://github.com/vllm-project/vllm-omni/pull/61**)
- **[feature] cpu worker (https://github.com/vllm-project/vllm-omni/pull/53)**
  - Mostly for the ease of local development
- [RFC] Support diffusers and community acceleration backends https://github.com/vllm-project/vllm-omni/issues/65

Meeting notes:

1. PR50, after verification on the local machine, it can be merged.
2. PR 61: Minor bug fix, ready to merge.
3. PR 53: Create device folder and move worker into it; add tests for multi-request streaming;
4. About NDA signature, Roger will help to negotiate with vLLM community.

Recording: Click here

2025-11-12 Meeting

Meeting: vLLM omni Meeting 2025-11-12 11:30 AM (UTC+8)
Join: Click here

Attendees:

- Gao Han (@Gaohan123, Huawei, wechat: gaohan046792)
- Liu Hongsheng (@hsliuustc0106, Huawei, wechat:)
- Huang Yongxiang (@SamitHuang, Huawei, wechat: samittech)
- Taichang Zhou (Bob Zhou , Huawei, wechat: ztc20010615)
- Roger Wang (@ywang96, vLLM, wechat: roger_99)
- Mo Zifeng (@lsotr0py, vLLM, wechat:)
- Jiangyun Zhu (@ZJY0516, vLLM, wechat: Z19984666173)
- WANG Cong (@congw729, Huawei, WeChat:hellocongw)

Agenda:

- Multi Request Streaming(https://github.com/vllm-project/vllm-omni/pull/51)
- Multimodal Inputs for Qwen(https://github.com/vllm-project/vllm-omni/pull/57)
- Documentation System Setup & Dependency Reorganization(#49)
- CI workflows & UT for omni_llm. (#50) Cong Wang

Meeting Notes:

1. PR 57: add config entry “require_multimodal_data” for all stages
2. PR 51: further improvements: the additional information and its update should be controlled by a well-defined interface, not a simple dict
3. PR 49: Roger Wang to look into how to make https://docs.vllm.ai/projects/vllm-omni/en/latest/ to work
4. Issue #56, Now all model implementations should be put into the repo of vllm-omni. Later when it is stable, we can consider migrating to vllm main repo.
5. PR 50 CI workflows & UT related:
    a. Essential but not urgent: align the setup environment steps with the document; remove the UT test files (code developers are in charge of their own code’s functionality before open sourcing)
    b. Urgent: Fix all the lint errors, and discuss with the corresponding developers if necessary.
    c. Roger Wong will confirm how to allocate & use the GPU resources for CI.
6. Issue #65 Support diffusers and community acceleration backends
    a. Can refer to verl’s config on sglang/vllm rollout backend, to improve the argument definition in {model_name}_{backend}.yaml, keep backend optimization argument naming the same as the original.
    b. The initial design fully depends on diffusers, we should consider taking vllm’s advantages on text encoding. Detailed interface to be discussed with @Isotr0py
    c. Acceleration backend support like fastvideo is not planned in v0.1 release

Recording: Click here