半导体行业策略：云巅千帆竞渡，端侧万物生辉，自主驭潮生

王竞萱 分析师
Email:wangjingxuan1@lczq.com
证书:S1320525020001

投资要点：

进入 2025 年以来，半导体行业利好频出，我们认为 2025 年在多重因素共振下行业有望实现进一步增长，迎来复苏的新阶段。在当前趋势下，从寻找各环节的最大公约数出发，我们看好以下方向：

云端 AI：算力需求增长依旧强劲，科技进步主体变化带来的边际影响或较为显著。预训练的算力膨胀方兴未艾，后训练和推理侧算力重要性愈发明显，成为维持行业快速增长斜率的关键动能，DeepSeek 的横空出世促使国产软硬件进一步结合，国产云端 AI 生态开辟新场景，有望实现对海外算力、模型、系统之间闭环关系的解耦，构建国产 AI 产业链的良性循环体系，建议关注：国产算力产业链。

端侧 AI：边缘 AI 设备的普及与端侧算力需求释放，或催生新的增长点。下一阶段的三大重点即：手机与 PC 市场的回暖，搭载端侧 AI 实现应用场景落地；AIoT 产品不断创新，以延伸人体关键感觉器官实现功能解放为目标；汽车领域的智能化普及趋势加速，全民智驾初步走入现实。整体来看端侧大模型发挥空间十足，看好具有增长确定性的细分赛道，建议关注：SoC、MCU、电源管理、智驾芯片、CIS 等。

自主可控：地缘政治与供应链重构背景下，中国半导体产业将加速技术突破与产能布局，同时全球产业链多元化趋势也将重塑行业竞争格局。消电国补如约而至，非 AI 领域有望逐渐走出底部区间，设备材料国产替代趋势未改，制造封测“Local for Local”有望受益，建议关注底部反转的可能性以及自主可控进程较快的子行业，包括：半导体设备、晶圆代工、先进封装等。

风险提示：

下游需求复苏不及预期；国产 AI 发展速度不及预期；地缘政治风险加剧

投资评级：看好（维持）

市场表现

相关报告

半导体 ETF：看周期趋势向好，多板块预示复苏

2024.07.09

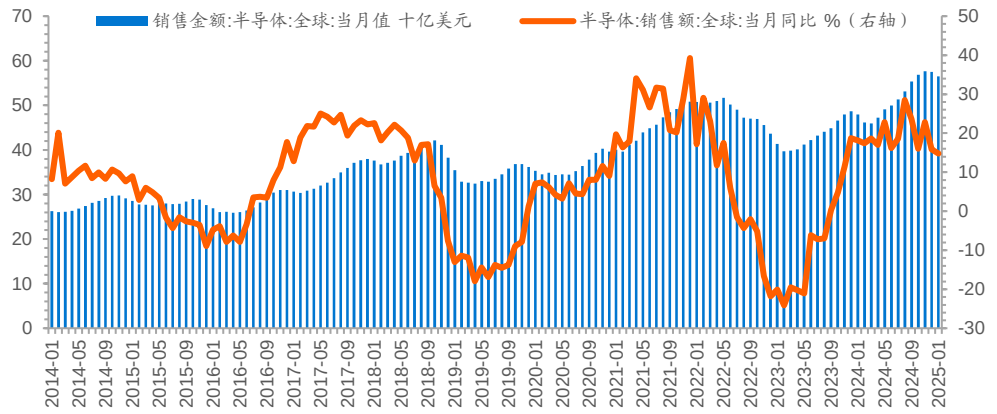
目 录

图表目录

1. 行情回顾

近两年半导体行业呈现“周期复苏+技术创新”双轮驱动格局：2024 年行业触底反弹，2024 年全球半导体总销售额达 6180 亿美元，同比增长 18.1%，2025 年预计突破 7000 亿美元。半导体本轮复苏主要由 AI 算力需求爆发、存储芯片价格反弹及国产替代加速推动。进入 2025 年，行业预计将进入复苏新阶段，AI 的云端和终端应用仍然是全球半导体行业的核心增长点，同时国产科技的爆发使得自主可控的趋势有望加速，政策层面的利好将贯穿本轮周期，整体来看可以期待半导体在成长和周期兼备的情况下有望进一步走向增长。

图1 半导体销售额及同比变化



资料来源：同花顺 iFinD，联储证券研究院

回顾近一年半导体行业市场行情，自 2024 年初至 2025 年 3 月 19 日，申万半导体指数涨幅达 40.21%，大幅跑赢沪深 300 的 18.42%，分阶段来看包含以下阶段：

①2024 年 1 月：由于消费电子需求疲软，叠加部分企业业绩预告不及预期，资金转向防守板块，行业成交额显著萎缩，半导体指数单边下跌 24.63%，创历史最大单月跌幅，但是同时单月半导体销售额延续 2023 年 12 月的亮眼走势，同增 18.15%，显示行业正在走出底部。

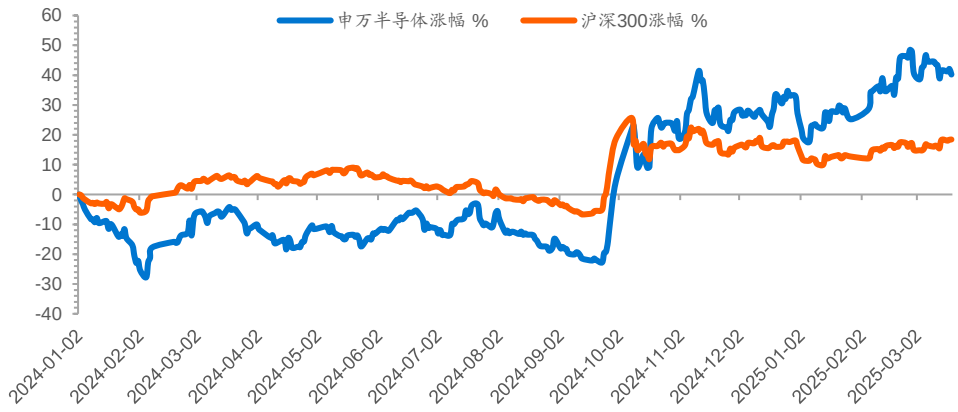
②2024 年 2-5 月：消费电子补库启动，智能手机、家电需求回暖；OpenAI 发布 Sora、英伟达 GB200 芯片发布等事件催化算力需求，带动国产云端算力芯片及服务器产业链走强；存储芯片库存情况得到改善，价格反弹，需求逐渐回暖；大基金三期成立，资金预期重点投向设备、材料等“卡脖子”环节，这一时期指数以震荡修复为主。

③2024 年 7-9 月：设备材料、晶圆制造等细分领域因国产替代加速，表现突出，H1 净利润增速超 100%企业占比达三成，复苏趋势明确，行业内呈现结构性分化，整体以积蓄力量的震荡表现为主。

④2024 年 10-12 月：以政策利好为主要推动力，财政货币政策宽松刺激市场信心，流动性得到大幅扩充，半导体板块领涨市场；地缘政治不确定性加速国产替代，海外限制事件频出，但同时国产技术在各环节加速发展，自主可控主线明确，市场表现为大幅上涨后高位震荡，期间最大涨幅达 83.10%。

⑤2025 年 1 月至今：2025 年开年，半导体行业呈现结构性分化与技术创新并行的特征，以 DeepSeek 为代表的一批科技企业在多个领域取得喜人成就，市场对于国产科技自立自强的预期拉升，硬科技板块持续获得资金流入，同时部分行业龙头业绩预增得到市场认可，年初至今行业指数涨幅达 10.21%。

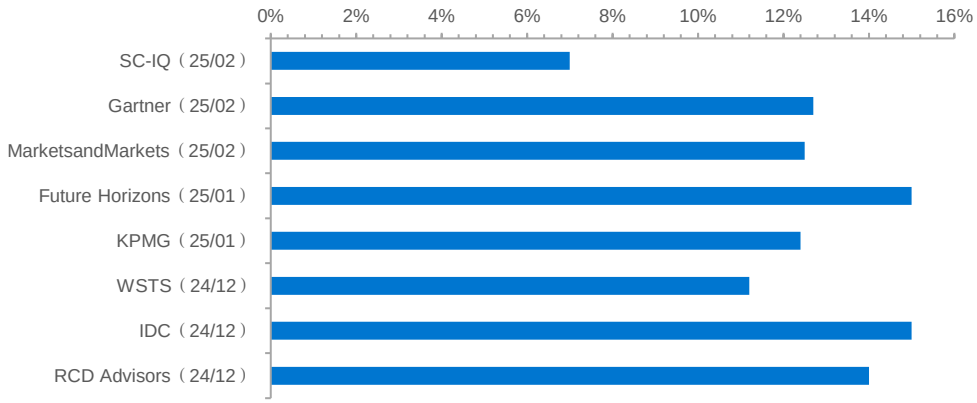
图2 近一年半导体行情复盘（截至 2025 年 3 月 19 日）



资料来源：同花顺 iFinD，联储证券研究院

行至 2025，**我们认为半导体行业整体乐观，行业进一步复苏的可能性较大。**据多家专业机构预测，2025 年半导体销售额增速将保持在 11%-15%之间，即基本延续 2024 年的增长速度。

图3 2025 年半导体销售额增速预测情况



资料来源：半导体产业纵横公众号，联储证券研究院

但是同时，我们认为行业的增长仍然并非是全面性的，而是以结构性的行情表现为主。从半导体行业发展历史来看，每一轮周期激发行业高斜率增长的核心因素在于关键的下游需求变化，而本轮周期中 AI 叙事的延续与否预计依旧是行业增长前景的最大扰动项。展望 2025，我们判断，“周期复苏+技术创新”的双轮驱动没有发生关键改变，寻找行业最大弹性的关键在于三个部分：第一，云端 AI 的算力逻辑；第二，AI 应用走向边缘端的可预期变化；第三，行业各环节自主可控的进程期待。

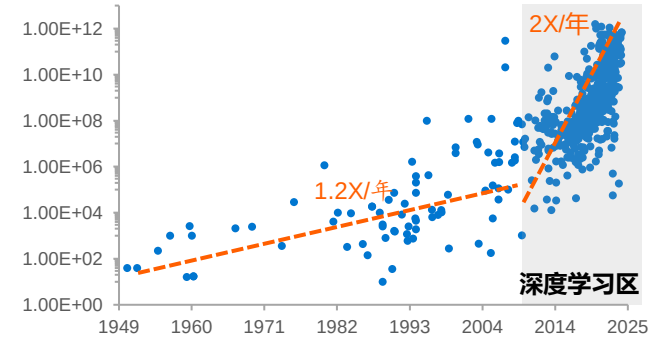
2. 云侧 AI：叙事逻辑无重大变化，叙事主体出现转向趋势

2.1 模型训练侧：Scaling Law 尚未见顶

Scaling Law 说明了 AI 模型训练领域的“大即是好”，是推动算力产业链爆发的核心法制。Scaling Law 最早由 OpenAI 于 2020 年提出，用以描述模型性能（损失值）与三个因素：模型的参数量、模型的算力投入和模型的训练 Token 数据集之间的幂律关系，即通过提升这三个因素而提升模型性能是有迹可循的。对于 Decorder-only 的模型而言，算力投入、参数量与训练数据量三者通常成线性正比关系，即随着模型参数量/训练数据量的提升，算力投入也需要相应地提升。

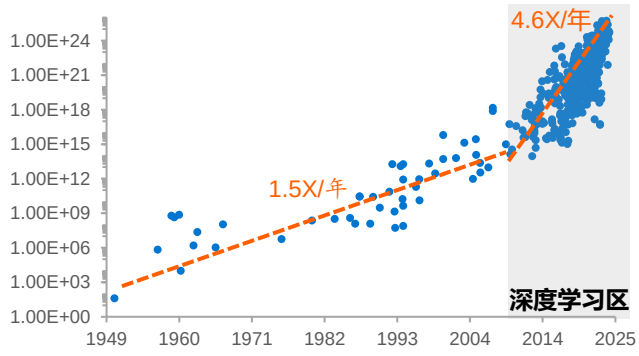
模型规模大小成指数上升，算力需求相应增长的规律未发生变化。纵观 AI 模型训练历史，自 2010 年后，新诞生的 AI 模型参数数量以平均每年 2 倍的速度提升，用于训练 AI 模型而投入的算力则以每年 4.6 倍速度提升。因此考虑模型的大小、训练数据量和计算量综合体现的模型规模大小始终处在快速增长中。因此我们认为，当前的新生 AI 模型仍然倾向于以提升模型规模大小的方式提升模型性能。

图4 大模型的参数数量持续增长



资料来源：Epoch AI，联储证券研究院
注：纵轴表示训练模型的参数数量

图5 大模型的算力投入持续增长

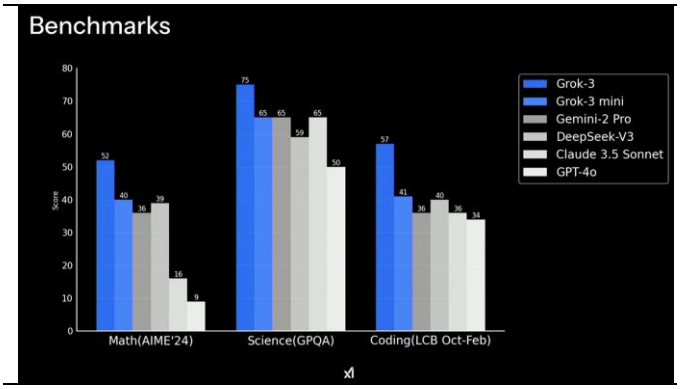


资料来源：：Epoch AI，联储证券研究院
注：纵轴表示训练模型的 Training Compute(FLOPs)

Grok-3 重磅发布，AI 模型表现持续推进。2025 年 2 月 18 日，Elon Musk 旗下的人工智能公司 xAI 发布了 Grok-3 系列模型，根据官方公开的测试结果，Grok-3 在包括 AIME 和 GPQA 等基准测试中超过了 DeepSeek-V3 和 GPT-4o 等顶尖模型，同时在大模型竞技场 Chatbot Arena 测试中，xAI 工程师表示，早期版本的 Grok-3 获得了第一的成绩，达到了 1402 分，成为全球第一个突破 1400 分的 AI 模型。

Grok-3 再次印证 Scaling Law，其出众表现离不开庞大的算力投入。xAI 第一阶段用了 122 天的时间构建了包含 10 万个 GPU 的数据中心 Colossus，创造了全世界规模最大的 H100 集群；但是团队认为此时的算力投入并未达到他们的需求，因此 xAI 又用了 92 天的时间将数据中心的规模扩展到 20 万个 GPU，而 Grok-3 的训练正是在这个基础设施上进行的。虽然 Grok-3 并未公布其模型的参数量，但是由训练其的算力投入倒推，预计 Grok-3 的参数量将远高于 314B 的 Grok-1。因此我们认为 Grok-3 的出现一方面既证实了当前 Scaling Law 仍然在发挥效用，AI 训练的“大力出奇迹”还在延续；另一方面，Grok-3 出色的性能表现会鼓励各 AI 研究机构继续加大算力投入以期获得出色的模型性能。

图6 Grok-3 在基准测试中的表现



资料来源：xAI Grok-3 发布会，联储证券研究院

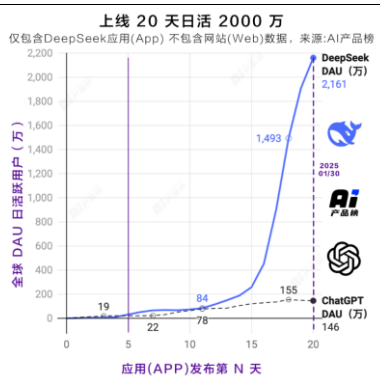
图7 Grok-3 在 xAI 数据中心 Colossus 训练得到

资料来源：xAI Grok-3 发布会，联储证券研究院

2.2 DeepSeek 引爆全球，聚焦模型推理能力

DeepSeek 一经发布迅速吸引全球目光。DeepSeek 于 2024 年 12 月 26 日开源 DeepSeek-V3 模型，2025 年 1 月 11 日发布 APP，1 月 20 日开源 DeepSeek-R1 模型。APP 上线 20 天全球日活 DAU 就突破了 2000 万，成为全球增速最快的 AI 应用；网站端访问量则以 22.3 倍的速度增长，1 月网站月访问量达 2.56 亿，环比增长 2230.89%；在 1 月累计获得 1.25 亿用户，其中 80%以上用户来自最后一周，即 DeepSeek 在没有任何广告投放的情况下，仅用了 7 天完成了 1 亿用户的增长。

图8 DeepSeek App DAU 迅速增长



资料来源：AI 产品榜公众号，联储证券研究院

图9 1 月份全球 AI 网站访问量排名

仅包含网站(web) 不包含应用(APP) 数据来源：AI产品榜

全球排名	AI产品榜	产品名 AI产品榜	网站(web)分类 aicpb.com	1月上榜网站 Web访问量	1月上榜网站 变化
1		ChatGPT	AI ChatBots	3.98B	4.31%
2		New Bing	AI Search Engine	1.89B	1.27%
3		Canva Text to Image	AI Design Tool	756.12M	0.89%
4		纳米AI搜索 | 原360AI搜	AI Search Engine	307.87M	-14.32%
5		Gemini	AI ChatBots	276.15M	2.38%
6		DeepSeek	AI ChatBots	256.54M	2230.89%
7		Character AI	AI Character Generat	228.18M	-1.45%
8		Deepl	AI Translate Tools	207.13M	4.18%
9		Notion AI	AI Writer Generator	164.54M	13.72%
10		Shop	E-COMMERCE	124.88M	-19.31%

资料来源:：AI 产品榜公众号，联储证券研究院

DeepSeek 的火爆表现主要源于其出色的推理表现，DeepSeek-R1 与 OpenAI-o1 的水平相当。DeepSeek-R1 的基准表现，R1 在多个评测中表现结果持平甚至超过了 OpenAI-o1，尤其是在数学领域的表现尤为突出，DeepSeek-R1 在 2024 年 AIME 上的单次预测准确率达到 79.8%；在 MATH-500 上，它取得了令人瞩目的 97.3 的分数，与 o1-1217 表现相当，显著超越了其他模型。在代码领域虽然表现差于 o1，但得分差距较小，在 Codeforces 上获得了 2029 Elo 评级，超过了 96.3%的竞赛参与者，整体来看，DeepSeek-R1 具有非常优秀的推理能力。

图10 DeepSeek-R1 的基准表现

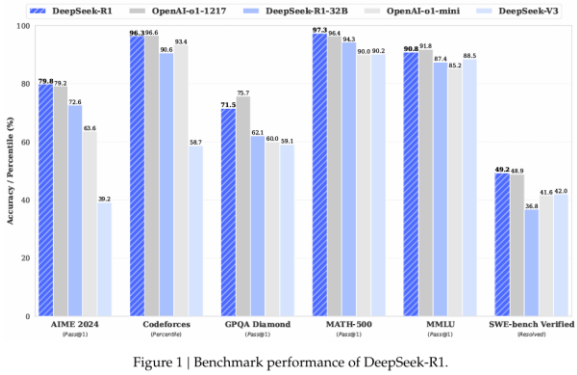


Figure 1 | Benchmark performance of DeepSeek-R1.

资料来源：DeepSeek 技术论文，联储证券研究院

图11 DeepSeek-R1 与其它代表模型的对比

	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

Table 4 | Comparison between DeepSeek-R1 and other representative models.

资料来源：DeepSeek 技术论文，联储证券研究院

DeepSeek-R1 其训练亮点在于探索了如何增强模型的推理能力。我们认为推理模型即具有回答复杂、多步骤、长链条问题能力的模型，其主要可用于解决如高级数学问题、高级编程问题等复杂问题。DeepSeek-R1 作为一个典型的推理模型，其推理能力的获得是通过以下三种方式实现的，并通过三种方式得到了三种模型：DeepSeek-R1-Zero、DeepSeek-R1、DeepSeek-R1-Distill:

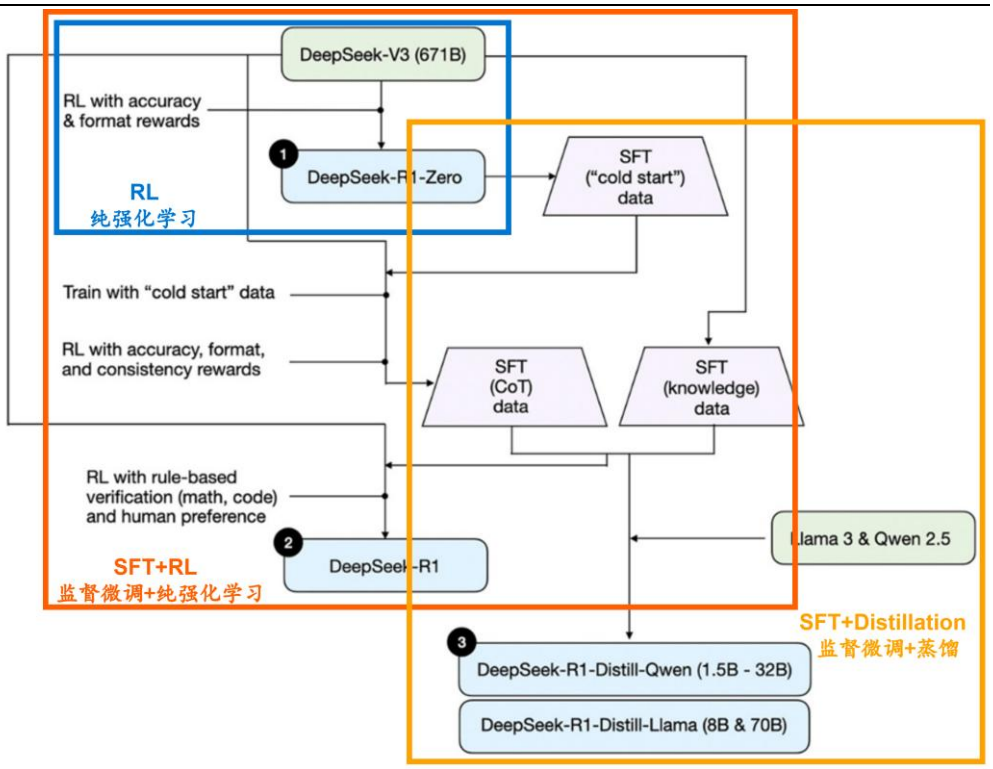
①RL（Reinforcement Learning，纯强化学习）：DeepSeek 团队设定了精度奖励和格式奖励两种奖励模型，精度奖励模型评估响应是否正确，格式奖励模型则要求模型提供思考过程，除此之外团队并未使用具有人类偏好的奖励模型，团队设计了一个简单的模板，引导模型首先生成推理过程，然后是最终答案，通过有意识地将约束限制在这

种结构格式上，训练得到了 R1-Zero 模型。

②RL+SFT（Supervised Fine-Tuning，监督微调）：团队使用未经过 SFT 的 R1-Zero 生成了冷启动 SFT 数据，进行了一轮微调，保留了精度奖励和格式奖励的同时，团队添加了一致性奖励模型，然后对模型开展了一个新的 RL 阶段，在这轮 RL 结束后的检查点团队又收集了 60 万个推理数据以及 20 万个由 V3 模型生成的 CoT（Chain of Thought，思维链）数据，并通过以上共 80 万个数据组成的数据集对模型进行了新一轮微调，在微调结束后，团队利用基于规则的奖励来指导数学、代码和逻辑推理领域，利用人类偏好奖励来指导通用数据，并进行了最后一轮 RL，最终得到了 R1 模型。

③SFT+蒸馏（Distillation）：为了让规模较小的模型也能获得类似于 R1 模型的推理能力，团队使用②中第二次微调所用到的 80 万个数据组成的数据集对 Qwen2.5-Math-1.5B、Qwen2.5-Math-7B 和 Llama-3.3-70B-Instruct 等小模型进行了微调，得到了若干蒸馏模型 R1-Distill。

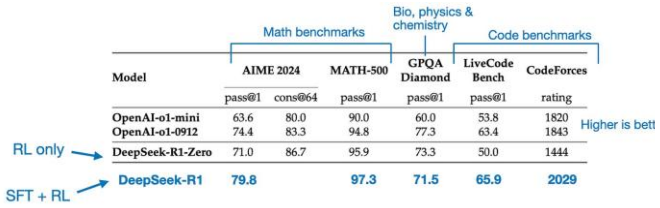
图12 DeepSeek-R1 的训练流程



资料来源：Ahead of AI，联储证券研究院

从效果观察，DeepSeek 团队的探索对于增强模型推理能力的意义巨大。对于仅进行了 RL 的 R1-Zero 模型，其推理能力在多个维度已经接近 OpenAI-o1，甚至于在 AIME（cons@64）以及 MATH-500 测试上的得分超过了 o1-0912，而在此基础上通过 SFT+RL 训练得到的 R1 的性能表现就更加出色了；对于蒸馏过后的模型，高效的 R1-7B 就能在所有指标上超越像 GPT-4o-0513 这样的非推理模型，R1-14B 在所有评估指标上都超过了 QwQ-32B-Preview，而 R1-32B 和 R1-70B 在大多数基准测试中显著超越了 o1-mini。

图13 RL 对增强模型推理能力的意义



资料来源：DeepSeek 技术论文，Ahead of AI，联储证券研究院

图14 蒸馏对增强模型推理能力的意义

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633
DeepSeek-R1-Zero	71.0		95.9	73.3	50.0	1444
DeepSeek-R1	79.8		97.3	71.5	65.9	2029

资料来源：DeepSeek 技术论文，Ahead of AI，联储证券研究院

DeepSeek 团队在 R1 模型的训练过程中的发现对于大模型推理能力的提升做出了诸多贡献：第一，直接将 RL 应用于基础模型，而不依赖 SFT 作为初步步骤，这种方法使模型能够探索 CoT 来解决复杂问题，从而开发出了展示自我验证、反思和生成长思维链等能力的 R1-Zero 模型，且这是第一次通过纯 RL 激发大语言模型推理能力的公开研究；第二，团队证明了较大模型的推理模式可以蒸馏到较小模型中，蒸馏模型虽然比不上 SFT+RL 得到的模型，但是却并不比纯 RL 模型表现差，而蒸馏模型的参数量是远远更小的；第三，蒸馏模型与在小模型上通过强化学习发现的推理模式相比，能带来更好的性能。

2.3 模型推理侧：当前推动算力需求增长的第二极

推理 Scaling Law 强调在推理阶段通过增加计算资源来提升模型性能。清华大学和卡内基梅隆联合研究团队展示了在不同推理策略（加权多数投票和 Best-of-N）下，GSM8k 数据集上的推理扩展情况。可以看到，随着每个问题的推理算力增加，测试误差率显著下降。这一现象表明，推理侧的 Scaling Law 不仅影响模型的计算效率，还直接决定了模型在实际应用中的性能表现。通过合理调整推理策略和模型规模，可以在有限的计算资源下，最大化模型的准确性和效率。因此我们认为，深入研究和优化推理侧的 Scaling Law，对于提升大规模模型的实际应用价值具有重要意义。

图15 加大推理的算力投入可以显著降低测试误差

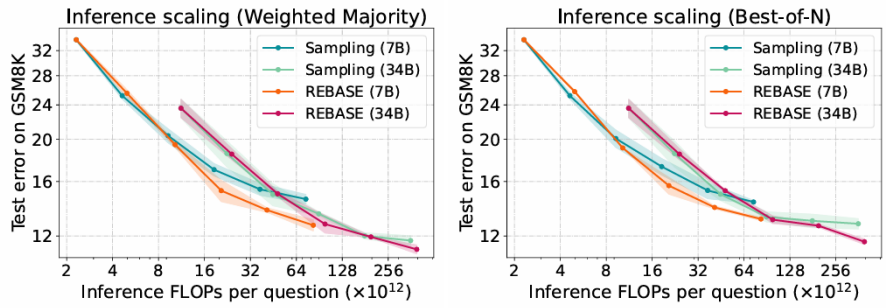


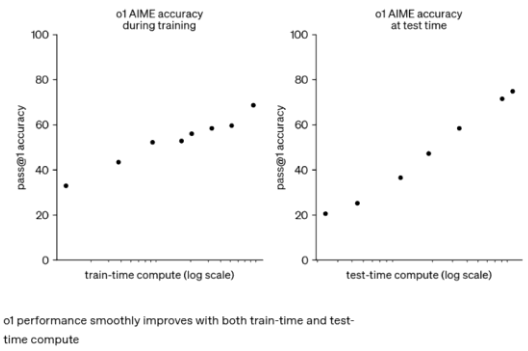
Figure 5: **GSM8k inference scaling across inference strategies and model sizes** (lower is better). The left/right panel shows the problem-solving error rate on GSM8K based on weighted majority/best-of-n. MCTS is not included in the comparison because of its poor compute-accuracy trade-off. REBASE is the compute-optimal inference strategy, and the optimal model size varies.

资料来源：Yangzhen Wu et al. (2024) Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models，联储证券研究院

随着模型规模的不断增大，推理侧的 Scaling Law 在提升模型性能方面的重要性愈发显著。OpenAI 团队通过大规模强化学习算法教会 o1 模型如何在高度数据高效的训练过程中使用 CoT 进行高效思考，无论是随着强化学习（train-time compute）的增加或

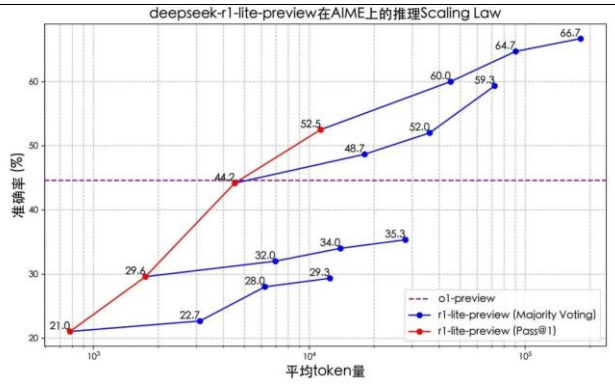
是思考时间的增加（test-time compute），o1 的性能均会不断提高。无独有偶，DeepSeek 团队发现 R1-Lite 模型在数学竞赛上的得分也与测试所允许思考的长度紧密相关，即 CoT 越长则推理结果越精准。因此我们认为，在当前 AI 发展阶段，在预训练阶段之外，在后训练阶段对模型加大强化学习力度或在推理阶段允许模型多思考一会儿，都能使模型的“智能”程度明显提升。

图16 o1 模型的表现与 train-time 和 test-time 均呈正比



资料来源：OpenAI 官网，联储证券研究院

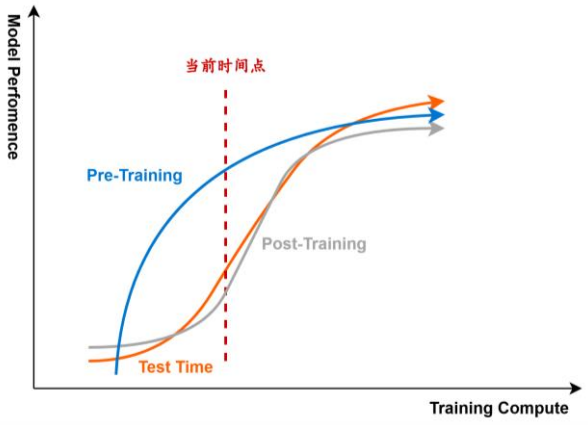
图17 DeepSeek R1 Lite 表现出的推理 Scaling Law



资料来源：DeepSeek 官微，联储证券研究院

后训练和推理的 Scaling Law 在当前时点下更可能有所作为。我们认为在当前时点：后训练阶段的 Scaling Law 强调在预训练完成后，通过 RL 和推理优化进一步提升模型性能，不仅能够提升模型的特定任务性能，还能在不显著增加模型参数的情况下实现性能飞跃；而推理阶段的 Scaling Law 则关注在模型部署阶段通过增加推理时间或计算资源来提升模型输出的质量。尽管当前的预训练阶段可能呈现出一定的参数扩展带来的边际收益逐渐递减的趋势，但是由于后训练和推理阶段的算力需求正接近“性能弹性”最大的阶段，因此发展 AI 模型整体的算力需求仍十分可观。

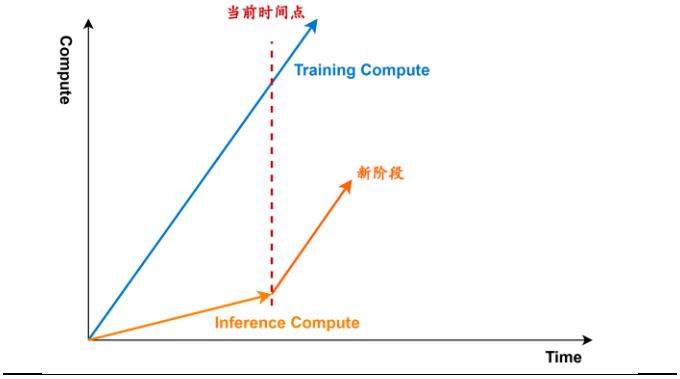
图18 Post-training 和 Inference 接棒 Scaling Law



资料来源：公开资料整理，联储证券研究院绘制

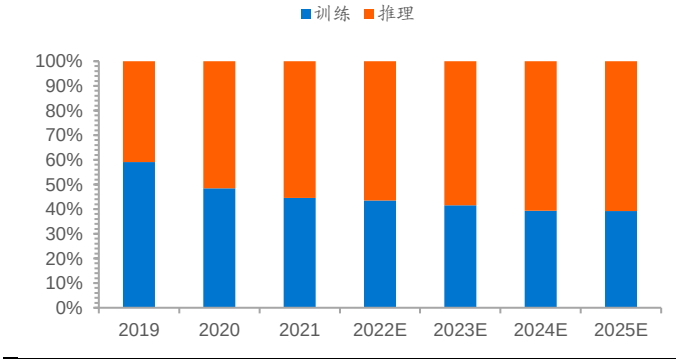
推理算力预计占比将不断提升，成为训练算力之外的第二极。我们认为随着 AI 的 2B/2C 端应用越发广泛，AI 使用将逐渐渗透到社会需求的各方面，对模型推理需求也势必会提升，叠加为追求模型性能优化而投入提升的 test-time，推理芯片将接棒训练芯片，有望复刻训练芯片的快速增长。根据 IDC 数据，2020 年中国数据中心用于推理的芯片的市场份额已经超过 50%，预计到 2025 年，用于推理的工作负载的芯片将达到 60.8%。预计单芯片的推理能力将逐渐增强，将单芯片算力耗尽的推理任务和小规模推理任务将出现混合部署趋势，芯片会逐步加强对于虚拟化技术的支持。

图19 推理算力预计将走向新阶段



资料来源：公开资料整理，联储证券研究院绘制

图20 推理芯片占比预计将进一步提升



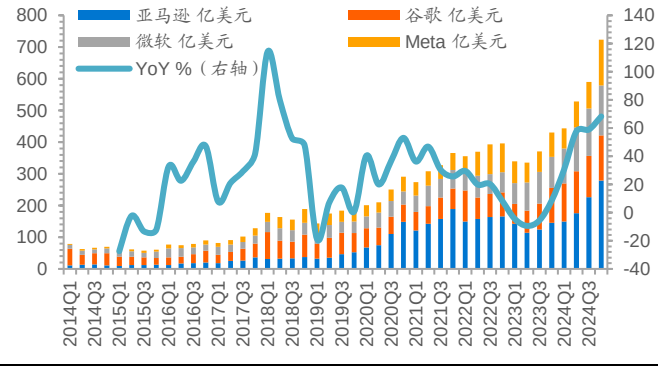
资料来源：IDC，联储证券研究院绘制

2.4 如何看待当前 AI 叙事逻辑下的算力需求？

四大 CSP 的 CapEx 依旧乐观，而 2025 年的支出可能达 3200 亿美元。北美四大 CSP 2024Q4 CapEx 合计达 723.48 亿美元，YoY 达 68.22%，而 2024 年全年合计 CapEx 达 2285.44 亿美元，YoY 达 54.87%。

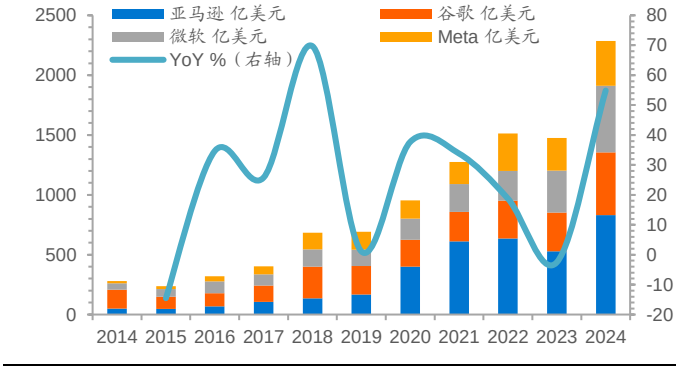
展望 2025，亚马逊表示公司 2025 年 CapEx 将增至 1000 亿美元；Meta 于电话会议中申明 2025 年 CapEx 将在 600 亿至 650 亿美元之间，这笔资金将用于推动 AI 战略；微软则是宣布，FA2025 将在 AI 数据中心方面开支 800 亿美元用于建设能够处理人工智能工作负载的数据中心；谷歌方面预期 2025 年资本开支 750 亿美元，YoY+43%。综合预计 2025 年北美四大 CSP 合计 CapEx 有望超过 3200 亿美元，按此估算 YoY 也在 40%以上。

图21 四大 CSP 季度 CapEx



资料来源：同花顺 iFinD，联储证券研究院
注：已将口径统一为日历年度

图22 四大 CSP 年度 CapEx



资料来源：同花顺 iFinD，联储证券研究院

DeepSeek 成本极低、算力有限实际上并不意味着全球算力需求萎缩。DeepSeek 通过高细粒度的 MoE（Mixture of Experts，混合专家）架构、MLA（Multi-head Latent Attention，多头潜在注意力）、FP8 混合精度训练、减少通信开销的 DualPipe 算法等算法层面的优化，实现了仅用 278.8 万个 H800 的 GPU 小时就训练出了 V3 模型，即按照 2 美元每小时租赁费用来算，训练成本仅为 557.6 万美元。与其它国际主流模型对比，DeepSeek 的训练成本极低，因此产生了部分认为未来 AI 研发算力需求将会大幅减少的观声音。

CSP 的乐观 CapEx 在一定程度打破了算力需求趋弱的观点。我们认为 CSP 持续加大投入的主要原因包括以下三方面：第一，DeepSeek 在技术层面上确实提供了降低模型训练成本的可能，但 AI 军备竞赛具有商业层面的考量，AI 技术领先可以在构建商

业生态时获得明显的先发优势；第二，即由于杰文斯悖论的存在，AI 算力需求仍然有较大的不降反升的可能，所谓杰文斯悖论即当技术进步提高了效率，资源消耗不仅没有减少，反而激增，例如，瓦特改良的蒸汽机让煤炭燃烧更加高效，但结果却是煤炭需求飙升，我们认为在 AI 领域也有可能会呈现这种发展趋势；第三，即我们在前面论证的，推理端算力成为下一阶段 AI 发展的重点，尤其对各大 CSP 而言，推理端实际上才是挂钩其收入的，因此模型推理效率的提升意味着算力投入的回报率提升，加大算力投入规模逻辑顺畅。

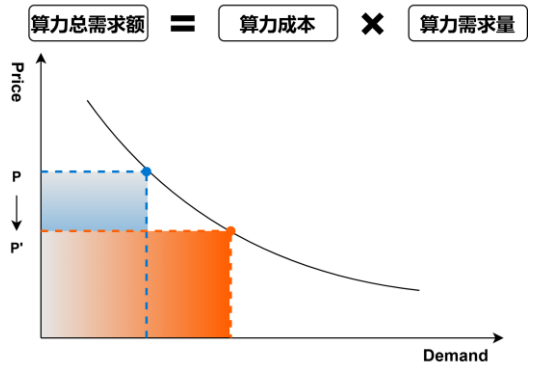
图23 DeepSeek-V3 的训练成本

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour

资料来源：DeepSeek 技术论文，联储证券研究院

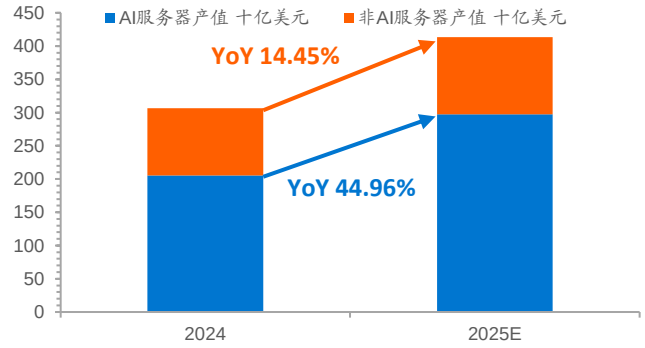
图24 杰文斯悖论使得总需求反而有望提升



资料来源：联储证券研究院绘制

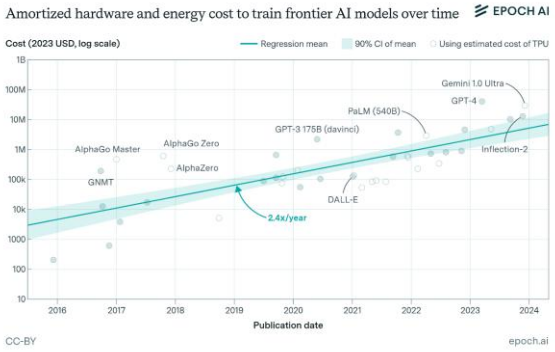
我们认为 AI 基建的叙事仍在延续，训练 AI 模型的成本或持续增长，AI 服务器需求仍将持续增长。从服务器看，根据 TrendForce 数据，2024 年整体服务器产值约 3060 亿美元，其中 AI 服务器较通用服务器增长动能更为强劲，产值约为 2050 亿美元，预计 2025 年 AI 服务器需求仍将持续增长，且价端有望提高较大贡献，产值有机会提升至近 2980 亿美元，YoY 达 44.96%，占整体服务器产值比例进一步提升至 7 成以上。从 AI 训练成本看，根据 Epoch AI 数据，从 2016 年到 2024 年，成本呈现出显著的上升趋势，前沿模型最终训练运行的摊销硬件和能源成本以每年 2.4 倍的速度快速增长，近年来的增长尤为明显，尽管技术取得了进步，但对更强大硬件和能源资源的需求继续推高成本，因此我们认为在不考虑算法优化的情况下，训练成本不断增长的趋势将持续演绎。

图25 2025 年服务器产值



资料来源：Trendforce，联储证券研究院

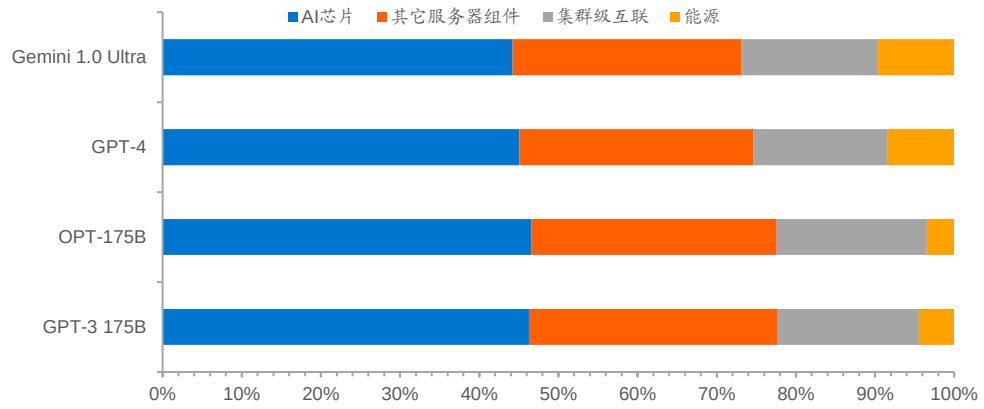
图26 训练 AI 模型的硬件和能源成本变化



资料来源：Epoch AI，联储证券研究院

不考虑人力成本的支出，芯片在 AI 的训练成本中占比最大，需求空间巨大。根据 Epoch AI 数据，芯片在 AI 训练成本中占据了显著的比例。以 Gemini 1.0 Ultra 和 GPT-4 为例，AI 芯片的成本占比均超过 40%，而在 OPT-175B 和 GPT-3 175B 中，这一比例更是接近 50%，这表明，随着模型规模的增大，AI 芯片在整体训练成本中的重要性愈发凸显。芯片作为计算的核心部件，直接决定了模型训练的效率和速度，因此 AI 研发团队对于高性能的芯片需求是天然巨大的，尽管其作为一次性支出的成本巨大，但是高性能芯片能够显著提升训练过程中的计算能力，从而缩短训练时间，降低单位计算成本。

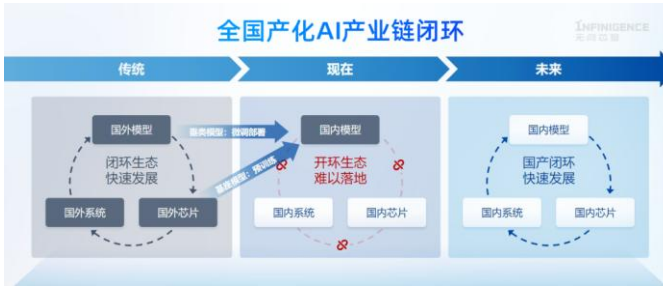
图27 训练 AI 模型的成本占比



资料来源：Epoch AI，联储证券研究院

国产算力产业链受益于 DeepSeek 发布，有望具有更强弹性。对于国内算力供应链而言，我们认为 DeepSeek 的出现是一剂强心针，标志着国产 AI 实现了比肩全球顶级模型水准的跨越，DeepSeek 通过算法、架构、工程的软硬件协同优化创新，以有限算力超低成本实现了性能比肩顶尖国外模型的国产模型，印证了软硬协同这一技术路线对推动 Scaling Law、突破算力瓶颈的有效性和巨大潜力。我们认为其对国产算力链的影响包括：第一，从当前国产 AI 生态来看，国产的算力资源仍然稀缺，随着以 DeepSeek 为代表的国产 AI 大模型的持续发展，国产算力缺口依旧较大，因此我们认为当前是软硬件协同实现国产算力芯片突破，进而实现国产 AI 生态闭环的关键窗口期，可以看到在 DeepSeek-R1 模型发布后国产硬件厂商用极快的速度纷纷接入 DeepSeek，我们判断通过此次软硬件合力发力的机会为二者未来持续融合夯实了基础；第二，如我们在前文的分析中指出推理芯片的占比将会持续提升，CSP 出于成本以及可得性考虑，我们判断 ASIC 在算力芯片中的比重将会不断提升，预计将持续为国产算力芯片创造发展动能。

图28 国产 AI 产业链有望实现闭环



资料来源：无问芯穹，智东西公众号，联储证券研究院

图29 国产芯片纷纷接入支持 DeepSeek

日期	企业	支持内容
0201	华为	华为云宣布与硅基流动联合首发并上线基于华为云昇腾云服务的 DeepSeek R1/V3 推理服务
0201	沐曦	Gitee AI 联合沐曦首发全国 DeepSeek R1 千问蒸馏模型，全免费体验
0204	天数智芯	成功完成与 DeepSeek R1 的适配工作，并且已正式上线多款大模型服务，其中包括 DeepSeek R1-Distill-Qwen-1.5B、DeepSeek R1-Distill-Qwen-7B、DeepSeek R1-Distill-Qwen-14B 等
0204	摩尔线程	基于 Ollama 开源框架，完成了对 DeepSeek-R1-Distill-Qwen-7B 蒸馏模型的部署，并在多种中文任务中展现了优异的性能
0204	海光信息	DeepSeek V3 和 R1 模型完成海光 DCU 适配并正式上线
0204	华为	并随原生鸿蒙创新推出基于昇腾算力的 DeepSeekR1 系列推理 API 及云链催服务
0205	沐曦	DeepSeek-V3 满血版在国产沐曦 GPU 首发推动上线
0205	华为	DeepSeek-R1、DeepSeek-V3、DeepSeek-V2、Janus-Pro 正式上线昇腾社区
0205	海光信息	海光 DCU 成功适配 DeepSeek-Janus-Pro 多模态大模型
0205	壁仞科技	DeepSeek R1 在壁仞国产 AI 算力平台发布，全系列模型一站式就能开发者创新
0205	太初元碁	基于太初 T100 加速卡 2 小时适配 DeepSeek-R1 系列模型，一键体验，免费 API 服务
0205	云天励飞	完成 DeepEdge10 "算力积木"芯片平台与 DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Llama-8B 大模型的适配，可以交付客户使用
0206	华为	华为 DCS AI 全栈解决方案中的重要产品—ModelEngine，全面支持 DeepSeek 大模型 R1&V3 和蒸馏系列模型的本地部署与优化，加速客户 AI 应用快速落地
0206	硅基科技	完成对 DeepSeek 全量模型的高效适配，包括 DeepSeek-R1/V3 671B 原生模型、DeepSeek-R1-Distill-Qwen-1.5B/7B/14B/32B、DeepSeek R1-Distill-Llama-8B/70B 等蒸馏模型，截至目前，DeepSeek 的全量模型已在庆阳、无锡、成都等智算中心完成了数万卡的快速部署
0206	昆仑芯	完成全版本模型适配，适用中包括 DeepSeek MoE 模型及其蒸馏的 Llama/Qwen 等小规模 dense 模型
0206	灵汐科技	完成了 DeepSeek-R1 系列模型在灵汐 KA200 芯片及相关智算平台适配，助力国产大模型与类脑智能硬件系统的深度融合
0206	燧原科技	全新一代的可重构数据流 AI 芯片 CAISA 430 成功适配 DeepSeek R1 蒸馏模型推理
0206	希姆计算	仅用数小时就将 DeepSeek-R1 全系列蒸馏模型快速适配到自研 RISC-V 开源指令集的推理加速卡系列之上，并落地全国多个千卡级以上智算中心
0207	算能	算能自研 RISC-V 开源指令集融合服务器 SRM1-20，成功适配并本地部署 DeepSeek-R1-Distill-Qwen-7B/1.5B 模型
0207	清微智能	可重构计算架构 RPU 芯片已完成 DeepSeek-R1 系列模型的适配和部署运行
0207	芯动力	芯动力神速适配 DeepSeek-R1 大模型,AI 芯片设计迈入"快车道"!
0207	亮芯中科	搭载亮芯 3 号 CPU 的设备成功启动运行 DeepSeek R1 7B 模型，实现本地化部署
0208	曦望	已完成 DeepSeek-V3 与 R1 全系列模型训推适配，率先可支持 V3 与 R1 671B 全量满血版模型部署

资料来源：智东西公众号，电子工程世界，联储证券研究院整理

综上，我们认为整个云端 AI 的算力需求扩张在 2025 年仍然值得期待。第一，pre-training 的 Scaling Law 仍在延续的背景下，post-training 和推理侧的 Scaling Law 接棒，各 AI 研究机构仍有足够动力加大各环节算力投入以追求模型性能的提升；第二，尽管 DeepSeek 的出现成功具象化了提升算法降低成本的路径，但是对各大 CSP 而言，AI 技术正面临商业化落地的关键窗口，加大投入即意味着提升模型推理能力进而实现云端服务收入落地；第三，DeepSeek 的出现一定程度上改善了之前仅有大型企业可以参与到 AI 前沿开发的竞争格局，DeepSeek 具有成本低廉、开源、性能强悍等特征，有望激发部分新生玩家的需求；第四，对于国产厂商而言，DeepSeek 成为了改变 AI 生态环境的先驱者，受其带动下国产硬件厂商切入参与到 AI 前沿产业链的机会增大。**因此我们建议持续关注国产算力产业链。**

3. 端侧 AI：模型推理能力提升，AI Agent 开启人机协作

我们认为在端侧的一个重点趋势即各类搭载 AI 功能的设备蓄势待发。其智能化表现有望得到大幅改善，2C 落地进程或加快实现，具体原因如下：

第一，DeepSeek 的发布使得在终端设备上部署 AI 的技术完备性得以提升。DeepSeek 模型在架构设计、训练优化及推理部署策略等多方面采用的创新技术，显著提升了模型性能和训练效率，降低了计算资源需求，为其在端侧设备部署 AI 带来了可能，其创新技术包括：

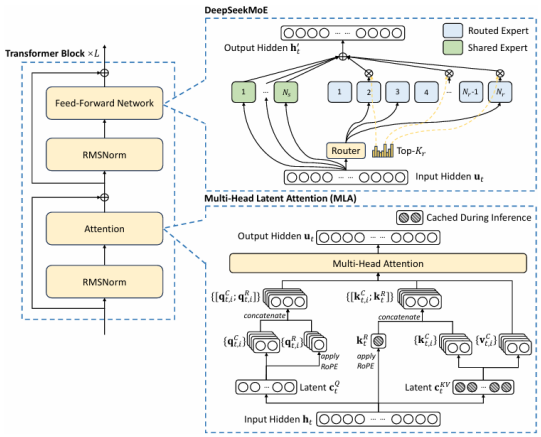
①DeepSeek MoE 架构：采用更细粒度的专家划分，并设置共享专家，提高了模型训练效率。同时，创新的无辅助损失负载均衡策略，通过动态调整偏差项确保专家负载平衡，避免了因负载不均衡导致的性能下降，使得模型在资源受限的端侧设备上也能稳定运行；

②MLA：通过对注意力键值进行低秩联合压缩，减少推理时的键值缓存，降低内存占用。仅需缓存特定向量，在保持性能的同时，显著减少了 Key Value 缓存，这对于端侧设备有限的内存资源来说至关重要，有助于模型在端侧设备上快速进行推理计算；

③推理阶段优化：在推理时，采用分离预填充和解码阶段的部署策略。预填充阶段通过合理配置张量并行、数据并行和专家并行，提高计算效率；解码阶段将共享专家视为路由专家，减少计算复杂度，利用 IBGDA 技术降低延迟提高通信效率，同时通过将两个微批次的相似计算工作量重叠提升了吞吐量并降低了全对全通信的开销，满足端侧设备对及时性的要求；

④模型蒸馏：前文我们提到的模型蒸馏通过将大型模型的知识转移到小型模型，使得在端侧设备有限的计算资源和存储条件下，依然能够实现高效的 AI 推理，提升了端侧 AI 应用的性能和用户体验。

图30 DeepSeek-V3 的架构



资料来源：DeepSeek 技术论文，联储证券研究院

第二，AI 模型发展迎来新范式，Agent 形式的 AI 在 2025 年迎来更多关注。AI Agent 是一种程序，它可以与环境互动，收集数据，并利用数据执行自决任务，以实现预定目标。人类设定目标，但 Agent 会独立选择实现这些目标所需的行动，Agent 的工作原理是将复杂的任务简单化和自动化。大多数自主 Agent 在执行指定的任务时，都会遵循特定的工作流程如下：

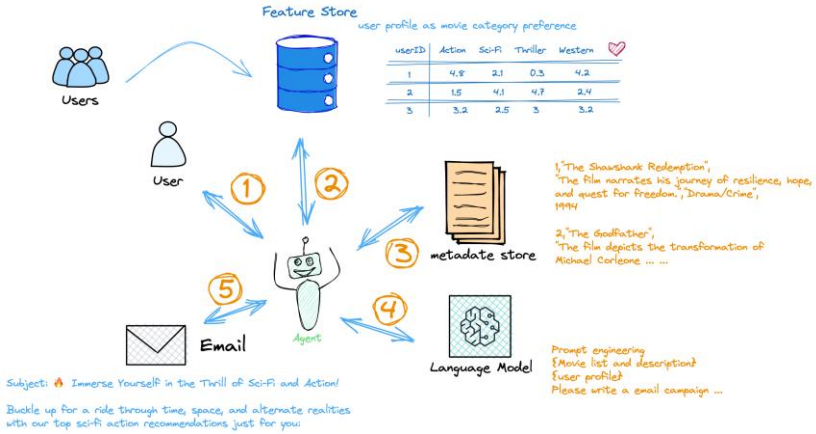
①确定目标：Agent 从用户那里接收特定的指令或目标，然后将目标分解成若干个

可执行的小任务，为了实现目标，Agent 会根据特定的指令或条件执行这些任务；

②获取信息：Agent 需要信息才能成功执行其计划的任务，因此 Agent 可能会访问互联网来搜索和检索所需的信息。在某些应用场景中，Agent 之间或 Agent 与其它模型交互，以获取或交换信息；

③执行任务：有了充足的数据，Agent 就会开始执行，完成一项任务后就将其从列表中删除，然后继续执行下一项任务，在完成任务的间隙，Agent 会通过寻求外部反馈和检查自身日志来评估是否达到了指定目标，同时在此过程中可能会创建并执行更多任务，以达到最终结果。

图31 AI Agent 的定义



资料来源：亚马逊官网，联储证券研究院

AI Agent、聊天机器人和生成式 AI 都是旨在帮助用户的 AI 形式。但是，它们在功能、复杂性和实际应用方面存在显著差异。功能性方面，Agent 是能够自主执行和适应各种任务的先进系统，旨在增强人类的能力，并且可以跨各个领域运行，而不仅仅是客户服务；复杂性方面，Agent 系统需要更复杂的技术理解上下文并有效地执行任务，由于可以从交互中学习并随着时间的推移而改进，因此它们通常适用于更复杂的应用程序；用户体验方面，Agent 通过处理多轮对话并根据用户行为和偏好提供个性化响应，可以以更自然、更像人类的方式学习和回应人类；投资成本方面，设置和运行 Agent 通常包括购买或开发 LLM、获取必要的硬件以及将系统集成到现有基础设施中，同时由于 Agent 需要大量高质量数据来训练和改善结果，因此额外的成本可能包括数据收集、存储和处理。

表1 AI Agent 和其它 AI 的区别

	AI Agents	非 Agent 聊天机器人	生成式 AI
功能性	自主行动和决策	处理线性响应且无适应性的简单对话	可以根据响应创建全新的内容
复杂性	使用 ML 和 NLP 等技术处理复杂任务	根据预设的规则进行行动，易于执行	使用本质上更复杂的深度学习技术
用户体验	根据用户行为的个性化和类人化反应	互动偏僵化（尤其是与偏离主题的问题）	高度互动的体验和动态对话
投资成本	需要更高的初始投资	通常更便宜且更易于部署	需要大量的初始投资

资料来源：TechTarget，联储证券研究院

从各大科技企业动作来看，AI Agent 在 2025 年或将大规模爆发。一方面，从技术角度看：AI Agent 正在变得更加多样化和智能化，Agent 在交互方式、性能提升、功能丰富等方面均有所提升；另一方面，从商业角度看：AI Agent 的商业化和生态建设正在加速，各大 CSP 均加快建设 Agent 与自身业务的沟通渠道，实现业务联动。因此我们

认为 2025 年 AI Agent 的变化，实际上是技术进步和市场需求共同作用的结果。技术上的发展，为 AI Agent 提供了强大的能力，而市场上的需求，以及用户对更智能、更便捷服务的追求，又推动了 AI Agent 的商业化和应用。

表2 近期各大科技企业在 AI Agent 方面的动作

时间	企业	内容
2024 年 6 月	腾讯	推出腾讯元器一站式 AI 智能体创作与发布平台，旨在帮助用户轻松创建和部署智能体，无需编写代码，即可实现聊天对话、内容创作、图像生成等功能。
2024 年 9 月	字节跳动	提出了基于强化学习的 LLM Agent 框架 ——AGILE，该框架下，Agent 能够拥有记忆、工具使用、规划、反思、与外界环境交互、主动求助专家等多种能力，并且通过强化学习实现所有能力的端到端训练。
2024 年 10 月	Anthropic	发布了 Claude 3.5 模型家族的更新，现有模型 Claude 3.5 Sonnet 获得了升级，获得了 Computer Use 能力，能够通过观看屏幕截图，实现移动光标、点击按钮、使用虚拟键盘输入文本等操作，真正模拟人类与计算机交互的方式。
2024 年 11 月	谷歌	谷歌云将提供从 AI Agent 的开发、部署到应用一站式商用生态，发布了 AI Agent 市场，用户可以在 AI Agent 市场中快速找到想要的 AI Agent，极大简化了客户的选择和部署流程。
2024 年 12 月	亚马逊	亚马逊宣布在旧金山成立新的研发实验室 Amazon AGI SF Lab，专注于构建 AI Agent 的基础能力，新实验室的主要目标是开发能够在数字和物理世界中执行行动的 AI Agent，使其能够处理复杂的工作流程，包括使用计算机、网页浏览器和代码解释器。
2024 年 12 月	微软	发布 Azure AI Agent Service，服务的一大特色是其强大的企业数据连接能力，支持数据基础化操作，包括与 Microsoft SharePoint 和 Microsoft Fabric 的无缝集成，以及工具联动以实现自动化操作。
2025 年 3 月	OpenAI	推出 Responses API，作为一个统一 API，可支持多轮交互和工具调用，同时推出网络搜索工具、文件搜索工具、计算机使用工具等多个工具，OpenAI 还预告未来一段时间内，计划发布更多工具和功能，进一步简化和加速在平台上构建智能体银行业。

资料来源：公开内容整理，联储证券研究院

第三，政策层面加速推动智能终端发展。2025 年政府工作报告中提出：持续推进“人工智能＋”行动，将数字技术与制造优势、市场优势更好结合起来，支持大模型广泛应用，大力发展智能网联新能源汽车、人工智能手机和电脑、智能机器人等新一代智能终端以及智能制造装备。

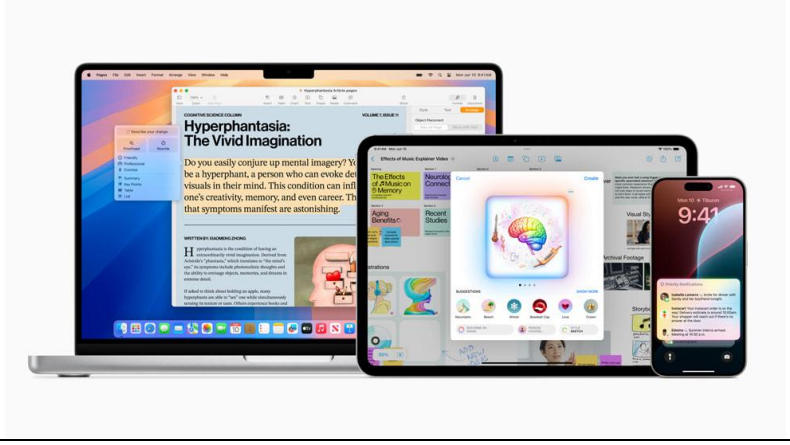
综上所述，我们认为边缘端 AI 在 2025 年将会迎来更多的发展机会，其中的重点细分赛道包括：AI 手机及 PC、AI 眼镜、汽车智能驾驶。

3.1 AI 手机&PC：端侧 AI 渗透的关键一年

苹果在 WWDC24 发布 Apple Intelligence，端侧 AI 进入新阶段。Apple Intelligence 深度集成于 iOS 18、iPadOS 18 和 macOS Sequoia 中，充分运用苹果芯片对语言和图像的理解与创作能力，可做出多种跨 app 操作，同时结合个人场景，为用户简化和加快日常任务流程。Apple Intelligence 主要具有以下能力：①理解和创作语言的能力，为用户解锁提高写作和沟通的新方式，包括对邮件、备忘录、Pages 文稿和各类第三方 app 中的文本内容的理解、改写和校对等；②借助于 Image Playground 为用户提供图像创作功能，帮助用户用全新方式进行交流和表达自我；③对于照片和视频可以进行便捷地搜索、修正和整理；④赋予 Siri 更深层次的语言理解能力，让 Siri 表现得更自然，更契合场景，更贴合用户个人需求，还能简化和加快日常任务流程等。

苹果 AI 或将于 2025 年支持中国大陆地区的苹果设备，或将带动 AI 手机逐渐成为标配。据苹果 CEO Tim Cook 在财报会议中表示，Apple Intelligence 将在 2025 年 4 月新增对包括中文在内的多种语言的支持。

图32 Apple Intelligence 登录 iPhone、iPad 和 Mac



资料来源：苹果官网，联储证券研究院

在 AI 手机供给方面，各手机厂商抢占 AI 高地，推动 AI 功能持续落地。各大手机厂商对手机的 AI 功能越发重视，积极尝试 AI 对各项功能的赋能作用，我们认为随着各品牌手机对 AI 功能的探索和理解加深，有助于推动手机的 AI 功能不断地深化扩展，同时厂商或将不断优化 AI 算法，提高 AI 功能的响应速度和准确性，进一步加大软硬件结合，实现更高效的数据处理和更流畅的用户体验。

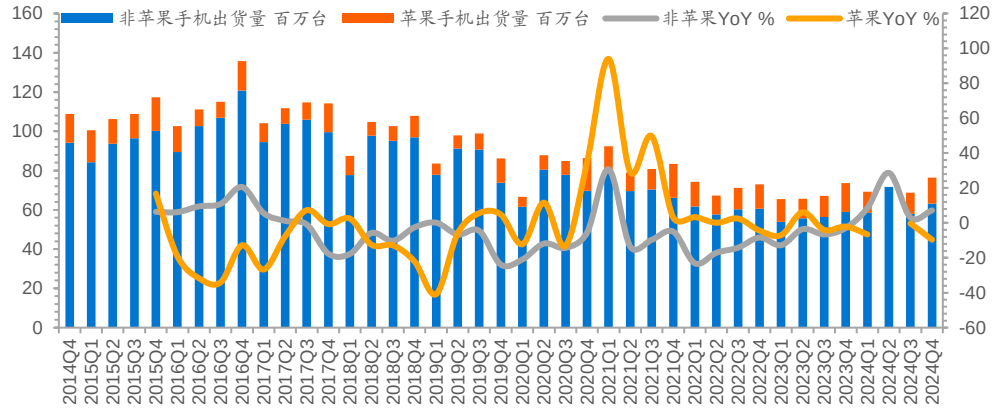
表3 近期各厂旗舰手机的 AI 功能

发布时间	品牌	型号	AI 功能
2024 年 10 月	vivo	X200	原子岛、写作助手、阅读助手、自动接听回复、通话录音总结和圈搜等
2024 年 10 月	OPPO	Find X8	一键问屏、相机问答、识屏问答、一圈即问、AI 千里长焦、图像助手和办公助手等
2024 年 10 月	小米	小米 15	超级小爱、AI 识屏、AI 写作、AI 识音、AI 字幕、AI 翻译和 AI 妙画等
2024 年 11 月	华为	mate70	运动轨迹、智控键、消息随身、主角时刻、隔空传送、降噪通话、时空穿越、通话摘要和静谧通话等
2025 年 2 月	三星	Galaxy S25	跨应用执行链、即时简报、即圈即搜、笔记助手、转录助手、通话助手、写作助手、照片助手、绘图助手和同传等

资料来源：各公司官网，联储证券研究院

在 AI 手机需求方面，用户换机时更考虑手机 AI 功能“有没有”而非“好不好”。苹果手机在中国地区出货量已连续 5 个月（可统计数据）同比下滑，2024Q4 中国地区苹果出货量 YoY 为-9.60%。与之相反，在中国地区的非苹果手机出货量连续 4 个月同比增加，24Q 出货量 YoY 达 7.18%。据苹果 CEO Tim Cook，苹果手机在华销量下滑的部分原因在于未能在中国地区提供 Apple Intelligence。

图33 中国地区苹果及非苹果手机出货量及同比变化



资料来源：IDC，同花顺 iFinD，联储证券研究院

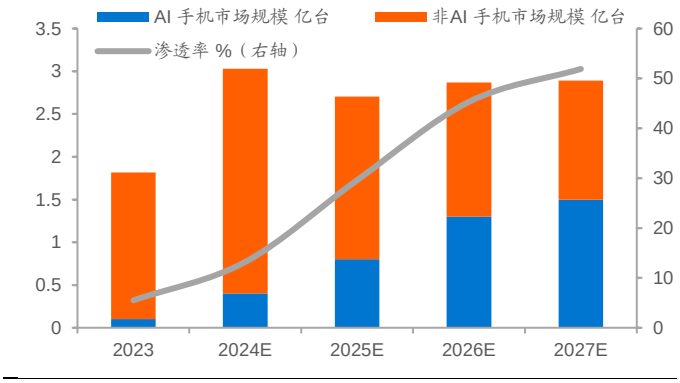
联想迅速推出部署端侧 DeepSeek 的 AI PC，有望助推 AI PC 市场吸引力。2025 年 2 月 25 日，联想集团发布 YOGA AIPC 新品，通过端侧部署与蒸馏技术创新，在消费级设备上实现 70 亿参数端侧 DeepSeek 模型的流畅运行。这一技术突破使得用户文档的总结、翻译、撰写等操作无需调用云端大模型即可完成，充分保障数据隐私与离线可用性，通过端云协同+混合式 AI 的架构创新，端侧部署 DeepSeek+个人云方案成功化解了大模型领域长期存在的“高性能、低成本、安全可靠”的“不可能三角”难题。我们认为联想作为先行者成功证实了现有 AI 模型在经过蒸馏后可以在端侧部署的可能性，或将带动更多设备尝试接入本地化 AI 功能。

图34 联想推出端侧部署 DeepSeek 的 AI PC

资料来源：联想，联储证券研究院

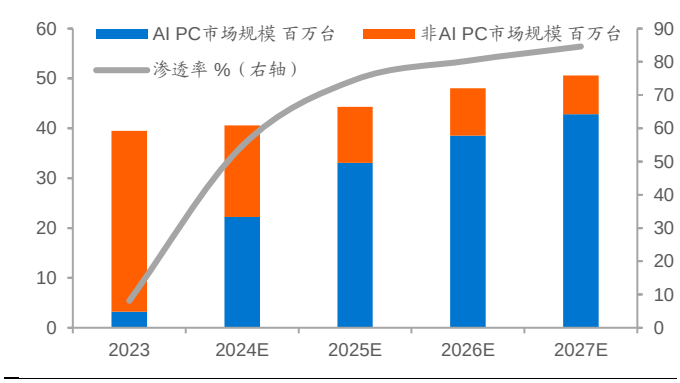
综合来看，AI 手机及 PC 在 2025 年或将进一步爆发，渗透率预计持续拉升。据 IDC 测算数据，2023 年我国 AI PC 的市场规模约为 320 万台，渗透率约为 8.1%；AI 手机的市场规模约为 1000 万台，渗透率约为 5.5%，还处于产品导入市场的阶段。而市场普遍认为 2024 年将是 AI 个人终端的爆发元年，由于 AI PC 具有生产力场景最多、个人算力最高、存储容量最大等优势，与 AI 技术融合速度会更快，2024 年市场规模有望达到 2200 万台，渗透率超过 50%，到 2027 年市场规模将突破 4200 万台；AI 手机囿于体积和功耗等限制，渗透速度相对较慢，但是智能手机的广泛性和普及性支持 AI 手机长期为智能机增长赋能，预计到 2027 年，AI 手机市场规模将超过 1.5 亿台，渗透率将超过 50%。

图35 AI 手机渗透率情况



资料来源：IDC，OPPO，联储证券研究院

图36 AI PC 渗透率情况



资料来源：IDC，联想，联储证券研究院

3.2 AIoT：以 AI 眼镜为代表的设备机会显现

跟进 Ray-Ban Meta，2024 年各厂商逐渐发力布局 AI 眼镜。Ray-Ban Meta 自 2023 年年底发布以来，便成为了现象级的爆款单品，甚至可以说成为了 AI 眼镜这一品类的专属代名词，我们认为 Ray-Ban Meta 的火爆具有多个原因：包括轻便、时尚、智能、社交等。而随着 Ray-Ban Meta 的火爆，可以看到在 2024 年出现了更多的对标产品，其中也不乏国产创业企业和互联网巨头，同时 Ray-Ban Meta 珠玉在前，科技厂商与传统眼镜厂商的抱团合作已成趋势，双方可在眼镜设计、光学配镜、市场渠道上产生深度合作，带来更大商业价值，伴随着 ODM/OEM 厂商的发力，供应链技术成熟度的提高，预计 2025 年将会有更多的 AI 眼镜，以及更多样化的单品设计。

表4 2024 年新发布的 AI 眼镜

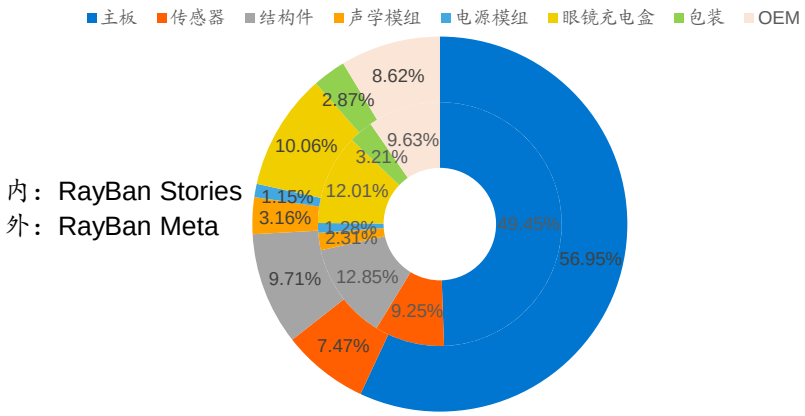
时间	产品	摄像头	售价	重量	其它卖点
2024 年 4 月	李未可 Meta Lens Chat	×	699 元	43g	拟人 AI 语音交互、最大 12h 续航、接入自研大模型 WAKE-AI
2024 年 5 月	闪极 AI 智能拍摄眼镜 A1	1600 万像素	1499 元	50g	展锐 AI 芯片、支持外挂存储和供电、LOHO 与科大讯飞合作
2024 年 6 月	Eddie Bauer	×	1684 元	不详	ChatGPT 交互、配备蓝牙、麦克风以及四声道扬声器
2024 年 7 月	Solos AirGo Vision	√	249 美元	34g（不含镜片）	多模态 A1、支持 GPT-40、可更换镜框设计
2024 年 8 月	界环 AI 眼镜(蜂巢科技)	×	699 元	41g	开放声场技术、AI 通知播报、面对面翻译
2024 年 10 月	Emteg Sense	√	待公布	62g	面部表情检测、情绪感知眼镜、记录食物消耗
2024 年 11 月	小度 AI 眼镜	1600 万像素	待公布	45g	AI 防抖算法、56 小时续航、中文大模型
2024 年 11 月	回车科技 Looktech	1300 万像素	199 美元	37g	声纹解锁、数码旋钮、智能体小程序
2024 年 12 月	加南 KANAAN-K1	√	1388 元	29.2g（不含镜片）	Live 图拍摄、AI 记忆胶囊

资料来源：VR 陀螺，联储证券研究院

以 Meta 眼镜为例看，成本压力有望得到缓解。RayBan Meta 成本总计 174 美元，主板芯片的成本约 99.1 美元，占比约 56.95%，成本占比超一半；眼镜充电盒的成本约 17.5 美元，占比约 10.06%；结构件的成本约 16.9 美元，占比约 9.71%；OEM 的成本约 15 美元，占比约 8.62%。我们认为参考 TWS 耳机发展历程，AI 眼镜有望演绎消费者购买单价下降，大幅拓展需求空间的趋势，形成如 TWS 耳机的“强需求”生态圈，原因包括：①技术成熟和供应链完善，主控芯片、传感器等核心元器件的技术逐渐成熟，

供应链也更加完善，使得生产成本降低，从而推动了价格的下降；②国产芯片的崛起，国产主控芯片的崛起使得 TWS 耳机的核心成本进一步降低，随着国产 AI 眼镜的发布，国产芯片预计持续导入；③制造良率的提升，生产效率提高，单位成本降低。

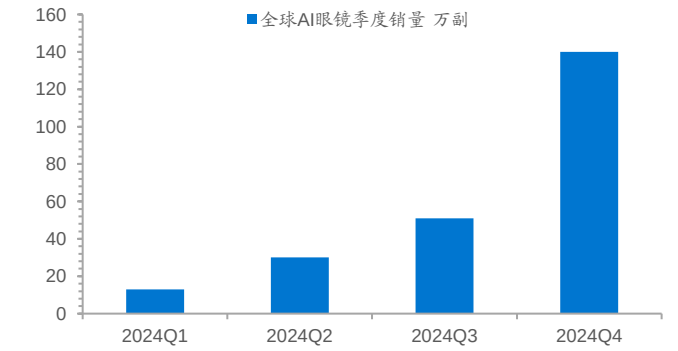
图37 Meta 两代眼镜 BOM 成本对比



资料来源：Wellsenn XR，联储证券研究院

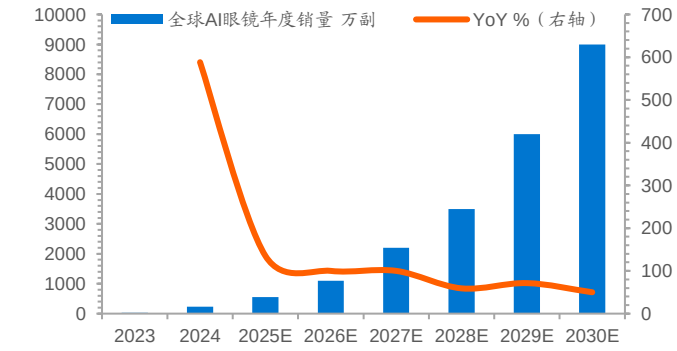
2025 年 AI 眼镜有望进入爆发期。我们认为眼镜作为人体感官的延伸，其便捷、高效、直接的特点，使得在眼镜上搭载端侧 AI 功能的意义性巨大，AI 眼镜未来的增长空间十分广阔。Ray-Ban Meta 的带动下，全球 2024 年的 AI 眼镜销量由 Q1 的 13 万副涨到了 Q4 的 140 万副，预计随着参与厂商的增多，这个数字还会以较快的速度持续增长，我们认为站在 2025 年初的节点上，可以看到多重因素驱动 AI 眼镜爆发，包括：Ray-Ban Meta 的销量持续增长、多款 A1 智能眼镜新品上市兑现、各大型企业入场发售 AI 智能眼镜新品。据 Wellsenn XR 数据预测，2025 年 AI 眼镜销量有望达到 350 万副，到 2030 年有望达 9000 万副，CAGR 达 83.7%。

图38 2024 年 AI 眼镜季度销量



资料来源：Wellsenn XR，联储证券研究院

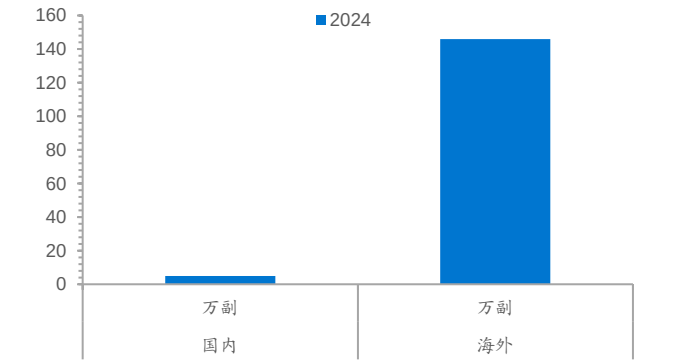
图39 AI 眼镜年度销量及预测



资料来源：Wellsenn XR，联储证券研究院

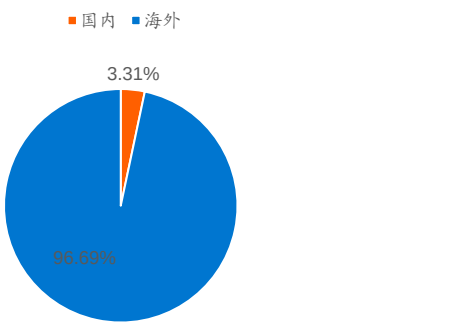
参考海外发展历程，国内 AI 眼镜处在爆发前夕。国内市场方面，2024 年国内销量仅有 5 万副，占比约 3%，销量主要来源于星际魅族 MYVU、StarV、界环、李未可等较早发布 AI 眼镜产品的厂商，四季度虽然发布产品较多，但量产发售的时间基本落在 2025 年，因此我们判断国内市场正处于 2023 年的海外市场阶段，即将迎来爆发。同时如百度、小米等互联网大厂 AI 眼镜单品曝光频繁，2025 年落地指日可待，有望形成“鲇鱼效应”，拉动消费者对 AI 眼镜的关注并推动消费者形成消费认知，充分提振国内市场。

图40 2024 年国内外 AI 眼镜销量（万副）



资料来源：Wellsenn XR，联储证券研究院

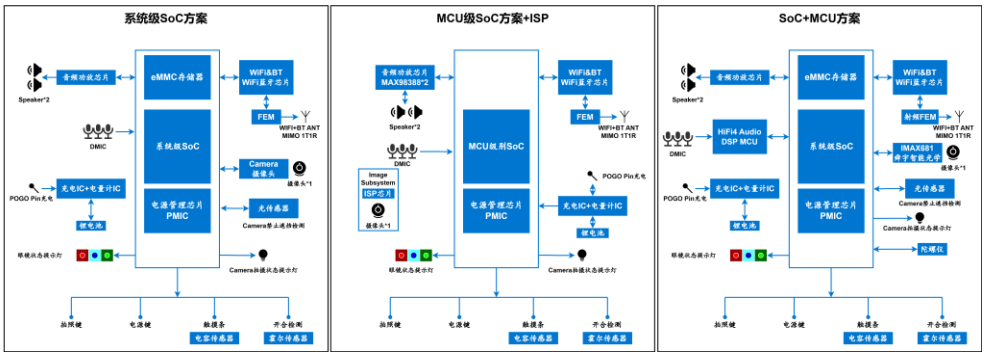
图41 2024 年国内外 AI 眼镜销量对比



资料来源：Wellsenn XR，联储证券研究院

带摄像头 AI 智能眼镜方案中，硬件的核心是 SoC。带摄像头 AI 智能眼镜目前有三种方案：系统级 SOC 方案、MCU 级 SOC+ISP 方案以及 SOC+MCU 方案，其中系统级 SOC 方案集成度较高，功能较多，内置支持拍摄功能的 ISP 模块。MCU 级别 SOC+ISP 方案集成度较低，需外接 ISP 芯片实现拍摄功能。SOC+MCU 方案适用性广，兼顾低功耗和高功耗应用，可通过系统调度有效控制续航时间。

图42 AI 眼镜的三种方案框架



资料来源：Wellsenn XR，联储证券研究院

①系统级 SOC 方案：SOC 中集成多核 CPU、GPU、DSP、ISP 等多种功能模块，集成度高，内核频率高，可为系统提供高性能计算能力，允许 AI 应用的端侧部署。同时集成的 DSP、ISP 可支持音频、摄影等功能，通用性高，方案成熟。缺点是成本高，功耗高。

②MCU 级 SOC+ISP 方案：MCU 级 SOC 内部以 MCU 内核为处理中心，可集成音频模块，GPU、NPU 或其他功能模块，集成度低，内核频率低，更多功能的实现需要外部连接其他功能芯片。优点是成本低，功耗低，定制化能力高。缺点是可提供的系统处理能力不高，且方案相对不成熟。

③SOC+MCU 方案：SOC 可负责需要高计算能力的应用场景，如支持分时操作系统，人工智能应用，拍摄功能等；MCU 可负责低计算能力的应用场景，如音频等。该方案下可实现合理的电源管理，延长设备运行时间，兼顾高低算力的应用需求，可用性广。缺点是成本极高，对芯片设计能力、系统开发能力要求高。

我们看好 AI 眼镜爆发对 SoC 的带动作用。考虑 AI 眼镜作为人体视觉感官的延伸，其应用场景落地于通过图像接入进行娱乐社交、生产力提升以及垂直领域的可能性更高，而初期需求依赖成本端的有效控制，随着使用习惯的培育未来或会对性能需求不断提升。

因此我们认为伴随 AI 眼镜爆发，SoC 器件受益可能性较高。

表5 AI 眼镜的三种方案对比

方案	SoC 方案	MCU+ISP 方案	SoC+MCU 方案
算力	高算力，支持 Linux、Android 等系统	低算力，支持 RTOS 等系统	高低算力兼备
AI 能力	支持，高 AI 能力	支持，低 AI 能力	支持，高 AI 能力
成本	高	低	极高
音频	支持	支持	支持
摄影	支持	支持	支持
连接方式	蓝牙、WiFi、esim	蓝牙、WiFi、esim	蓝牙、WiFi、esim

资料来源：Wellsenn XR，联储证券研究院

3.3 智能驾驶：智驾逐渐成为亲民标配，市场空间有望大幅扩容

比亚迪智驾全面下放，天神之眼覆盖大部分车型，汽车价格战或将由“电动战”转向“智能战”。2 月 10 日，比亚迪举办智能化战略发布会，宣布比亚迪将全系搭载“天神之眼”高阶智驾系统，开启“全民智驾时代”，比亚迪全系自研的天神之眼基本覆盖了旗下的 90%以上的车型，三套方案 A/B/C 分别对应高端子品牌仰望/中高端腾势+比亚迪/低价位段车型。面向大众主流市场的天神之眼 C。由于将高快领航、城区记忆领航等中阶智驾系统的搭载门槛降至了 10 万以内，搭载天神之眼 C 的海鸥智驾版起售价仅为 6.98 万元，而比亚迪本身在中低端市场较大的销售体量与市场认可度，我们认为比亚迪此举有望对智能驾驶现有市场格局产生较大改变。

图43 比亚迪天神之眼首批上市车型价格

资料来源：比亚迪官微，联储证券研究院

特斯拉 FSD 入华，智驾竞争再次深化。2 月 25 日，特斯拉中国发布资讯，表示开始面向定购了 FSD 智能辅助驾驶功能的用户，分批推送软件更新，升级城市道路 Autopilot 自动辅助驾驶。特斯拉在此次软件更新中最主要的更新为对车辆提供自动辅助驾驶，会根据导航路线引导车辆驶出匝道和交叉口，在路口识别交通信号灯进行直行，左转，右转，掉头等，并根据速度和路线自动进行变道动作。此次更新虽然并非完整的全自动驾驶 FSD，但是 Autopilot 的更新使得特斯拉在中国的高阶智驾方面追平国产厂商。作为全球新能源汽车的标志性企业，特斯拉此举或将进一步刺激国内智驾市场，供需双方强化聚焦当前智驾的配备情况。

图44 特斯拉为国内用户提供 Autopilot

2024.45.32.12软件更新

特斯拉资讯 2025-02-25

#软件更新 #自动辅助驾驶 #智能驾驶 #OTA

2024.45.32.12已开始分批次推送，本次软件更新主要升级内容为：

1.城市道路Autopilot自动辅助驾驶（优化现有NOA自动辅助导航驾驶功能）：在通行受控道路（道路使用者通过匝道入口和匝道出口进出的主干道）和城市道路上使用Autopilot自动辅助驾驶，会根据导航路线引导车辆驶出匝道和交叉口，在路口识别交通信号灯进行直行，左转，右转，掉头等动作。并根据速度和路线自动进行变道动作。在不设置导航路线时，会根据道路实际情况选择最优道路行驶。

2.驾驶室摄像头：您后视镜上方的驾驶舱摄像头现在可以判断驾驶员的注意力是否集中，并通过警报，提醒您在智能辅助驾驶系统启动时将注意力集中在道路上。驾驶室摄像头视频在车辆内部进行处理。任何人（包括Tesla公司）均无权访问。

3.地图包版本更新：CN-2025.8-15218。

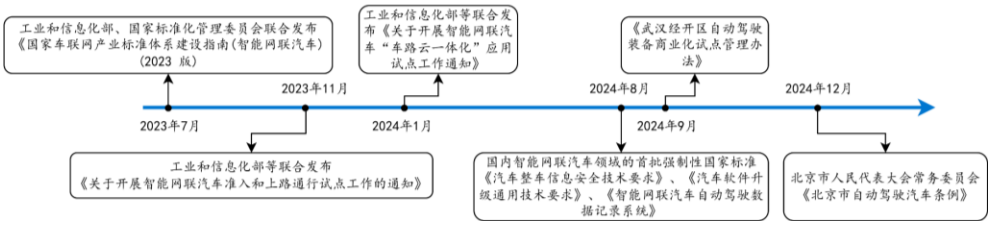
*部分功能实现时间和效果可能根据车型和车辆配置存在差异。

受控道路和城市道路Autopilot自动辅助驾驶功能已在部分车型上推出，并将逐步扩展适配的车型范围。如您已购买上述功能，需了解您的车辆适配情况，可通过特斯拉App消息中心联系"在线客服"或拨打400客服热线查询。

资料来源：特斯拉官微，联储证券研究院

政策层面对智驾的支持力度逐渐加大。从政策面看，各部门对于智驾的认可及支持都在不断加深。以北京市最新政策条例看：北京市第十六届人民代表大会常务委员会第十四次会议通过了北京市自动驾驶汽车条例，在基础设施规划建设、上路通行管理、安全保障、法律责任等方面都制定了相关管理办法，此条例将从 2025 年 4 月 1 日开始实行，我们认为随着顶层的政策构建愈发完整，智驾的渗透率和使用率不断提升的可能性也就越高。

图45 智驾政策逐渐放开



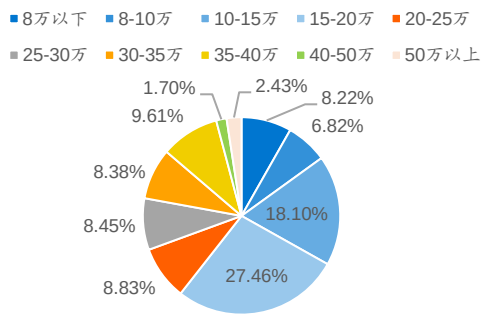
资料来源：各部门，联储证券研究院绘制

从 2024 年 20-25 万区间来看，20 万以下车型高阶智驾 2025 年有望爆发。高阶智驾功能的定价格局在三年间经历了显著变迁。2022 年，该技术主要集中于 30-35 万元的中高端车型，被视为豪华的象征。进入 2023 年后，市场呈现阶梯式分化：一方面向 50 万元以上的旗舰车型延伸以强化技术标杆地位，另一方面快速下沉至 20-25 万元主流区间，当年下半年该价格段成为城市 NOA 功能量产的核心战场。而进入 2024 年后，高阶智驾一方面持续下沉，另一方面则加速渗透：

根据中汽协数据，以销量统计，2024 年我国售价在 20 万元以下的新能源汽车占比较高，合计达到 60.60%，即对于消费者而言 20 万元以下车型为主力购买区间。根据佐思汽研数据，2024 年 1 月，20-25 万元价位段汽车的城市 NOA 搭载率仅有 2.1%，至 2024 年 10 月，该值则升至 24.7%，这一变化标志着城市 NOA 正加速普及，20-25 万价位段已成为车企及第三方智驾厂商竞争的前沿阵地。

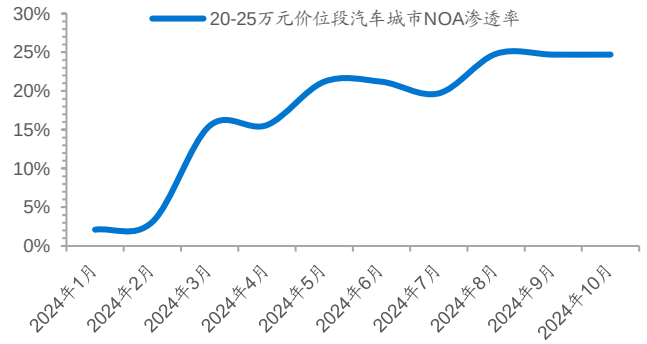
伴随各大车企及第三方智驾厂商在智驾配备和智驾技术上的持续发力，我们预计 2025 年搭配高阶智驾的低价位段车型会进一步提升。而由于此部分车型具有较大的市场份额，因此预计智驾市场有望迎来爆发。

图46 2024 年新能源汽车价位段占比



资料来源：中汽协，同花顺 iFinD，联储证券研究院

图47 2024 年 20-25 万元价格区间内的城市 NOA 汽车渗透率



资料来源：佐思汽研，联储证券研究院

智驾走向平权时代，我们认为核心受益方包括国产智驾芯片及 CIS 等厂商。随着智驾向各价位段车型的普及，我们预计智驾行业或将出现以下两大趋势：第一，DeepSeek 的开源推广，有望支持在既定算力有限情况下的模型性能提升，端到端大模型性能提升使得国产汽车算力芯片性价比突出，发挥空间扩容，可以用较低成本应用于低价位段车型；第二，视觉方案相较于含激光雷达的多传感器方案，具有更低的成本，以比亚迪三套天神之眼配备硬件来看，C 方案在不含激光雷达的情况下，选择加装了更多的摄像头，因此我们判断在智驾快速普及时期，采用视觉方案的车型会进一步提升，对作为摄像头核心器件的 CIS 而言是较大利好。

表6 比亚迪三套天神之眼对比

	域控成本	芯片	摄像头	前视摄像头	毫米波雷达	激光雷达	超声波雷达	适用车型
天神之眼 A	预估 8000-12000 元	2×Orin X	11 个	2 个	5 个	3 个	12 个	仰望
天神之眼 B	预估 5000-7000 元	Orin X	11/12 个	2/3 个	5 个	1/2 个	12 个	腾势+比亚迪
天神之眼 C	预估 3000-4000 元	Orin N（70%）/地平线 J6（30%）	12 个	3 个	5 个	×	12 个	比亚迪

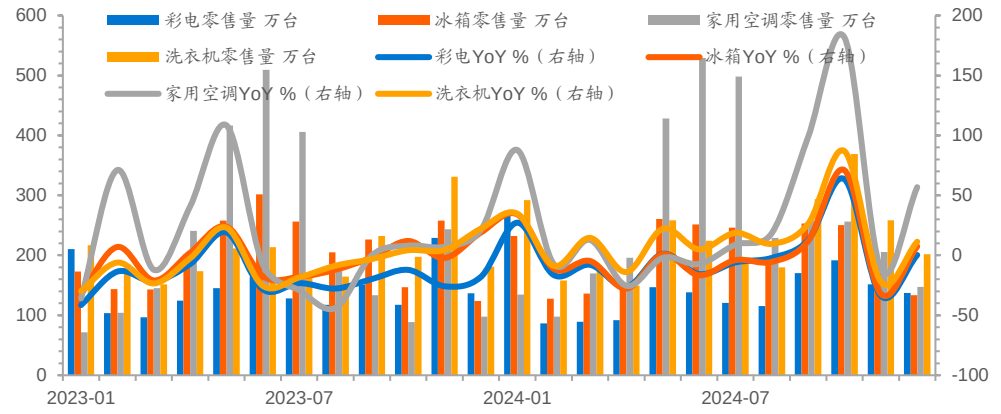
资料来源：比亚迪官网，芝能汽车，联储证券研究院

4. 自主可控：长期坚定不移地看好科技自立

4.1 消费电子：国补弹性测度，有望带动产业链新活力

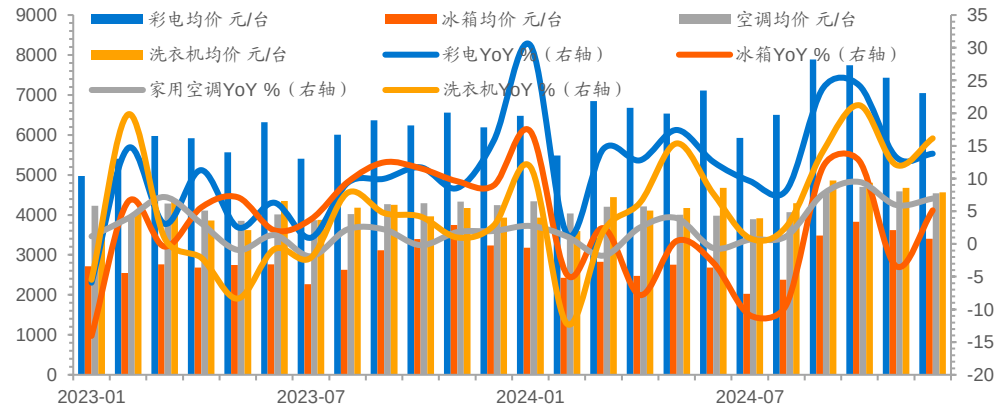
家电国补是推动 2024 年家电行业复苏的重要推手。2024 年 8 月 24 日，商务部等 4 部门办公厅发布关于进一步做好家电以旧换新工作的通知，对个人消费者购买 2 级及以上能效或水效标准的冰箱、洗衣机、电视、空调、电脑、热水器、家用灶具、吸油烟机 8 类家电产品给予以旧换新补贴。以彩电、冰箱、空调、洗衣机四大件来看，自国补政策出台后，其零售量和均价都出现了不同程度的提升，尤其是在 9、10 月份，同比增长值均出现了“小高峰”。

图48 主要家电品类零售量及同比变化



资料来源：产业在线，同花顺 iFinD，联储证券研究院

图49 主要家电品类均价及同比变化



资料来源：产业在线，同花顺 iFinD，联储证券研究院

家电国补对于家电的零售量及均价都带来了正向影响。从 2024 年 9 月划分前后两个阶段，分别称为“国补前”和“国补后”，从零售量来看：家电四大件在国补后的增长显著高于国补前阶段的增长，增长值至少为 5.8pct，其中空调增长值最多，达到了 43.5pct；从销售均价来看：家电四大件同样出现了国补后的明显提升，最低的增长值出现在彩电为 5.6pct，最高的则为洗衣机，达 11.8pct。因此可以明显看到，家电国补极大地提升了消费者热情，对于家电以旧换新的充分需求使得家电销量实现了一定增长，同时，由于补贴的存在，对于消费者而言，既定额的消费支出可以购得相对高端的产品，国补前原本零售价格较高的产品的到手价也有一定降幅，因此消费更高价格产品的倾向明显。

表7 补贴前后家电零售量变化情况

	彩电零售量（万台）	冰箱零售量（万台）	空调零售量（万台）	洗衣机零售量（万台）	彩电 YoY	冰箱 YoY	家用空调 YoY	洗衣机 YoY
2023 年 1-8 月	1090.24	1682.68	2069.14	1467.64				
2023 年 9-12 月	634.17	754.74	563.70	942.23				
2024 年 1-8 月	1056.07	1593.88	2282.20	1629.28	-3.13%	-5.28%	10.30%	11.01%
2024 年 9-12 月	650.89	812.18	867.13	1122.54	2.64%	7.61%	53.83%	19.14%

资料来源：产业在线，同花顺 iFinD，联储证券研究院

表8 补贴前后家电均价变化情况

	彩电均价（元/台）	冰箱均价（元/台）	空调均价（元/台）	洗衣机均价（元/台）	彩电 YoY	冰箱 YoY	家用空调 YoY	洗衣机 YoY
2023 年 1-8 月	5695.36	2637.69	4044.43	3982.49				
2023 年 9-12 月	6339.13	3374.30	4283.22	4080.90				
2024 年 1-8 月	6444.69	2593.63	4093.21	4142.61	13.16%	-1.67%	1.21%	4.02%
2024 年 9-12 月	7527.14	3583.27	4604.58	4726.43	18.74%	6.19%	7.50%	15.82%

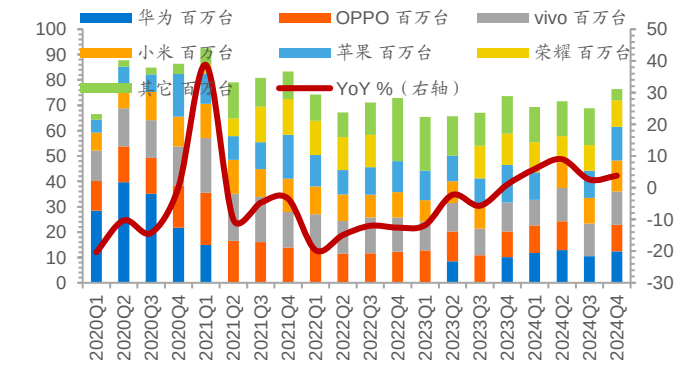
资料来源：产业在线，同花顺 iFinD，联储证券研究院

从家电国补推演手机国补影响，有望拉动国内手机需求大涨。2025 年 1 月 5 日，发改委、财政部发布关于 2025 年加力扩围实施大规模设备更新和消费品以旧换新政策的通知，实施手机等数码产品购新补贴。对个人消费者购买手机、平板、智能手表手环等 3 类数码产品（单件销售价格不超过 6000 元），按产品销售价格的 15%给予补贴，每位消费者每类产品可补贴 1 件，每件补贴不超过 500 元。

2024 年全年国内智能机出货量 YoY 为 5.26%，以此作为假设：国内手机市场的当前自然增长率为约 5%，考虑我国手机价位段占比情况，消费者为充分利用补贴机会，中高价位段且位于补贴范围内的手机需求增长的可能性更高，我们认为销量提升最大的部分为 3000-4000 价位段，其次是 4000-6000 价位段，再次是 3000 以下，最后是 6000 以上。

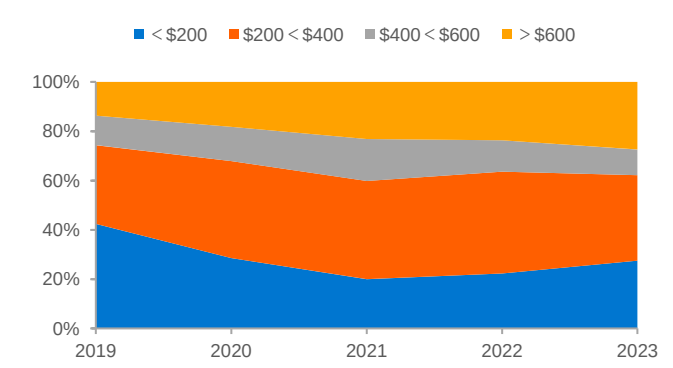
参考家电补贴后的影响，估计增幅最大部分的价位段同增提升 8-10pct，增幅最小部分的价位段同增提升 4-5pct，同时考虑到国补带来的促销作用或会随时间有一定减弱，综合预计 2025H1 智能机同比增长有望达到 10.5%-11.7%，H2 同比增长也有望达 5.5%-8.2%。

图50 中国地区智能手机出货情况



资料来源：IDC，同花顺 iFinD，联储证券研究院

图51 中国智能手机价位段



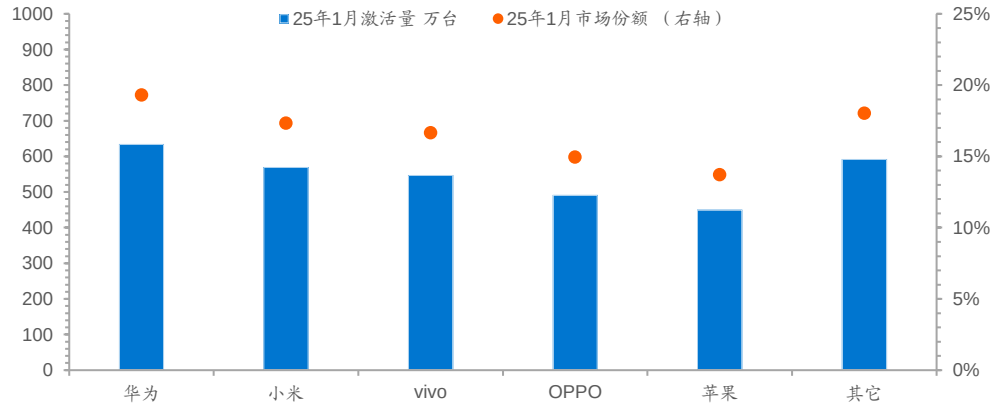
资料来源：IDC，联储证券研究院

手机国补助力国内手机市场增长，国产手机厂商或为主要受益方。原因在于：第一，2025 年 1 月，华为以 19.3%的市场份额位居榜首，激活量达 634.81 万台，同比增长 27.7%，小米、vivo 则分别以 17.34%和 16.66%的份额紧随其后，苹果仅有 13.71%的份额位列第五，显示国产头部品牌已形成一定优势；第二，国产品牌在中低端机型市场具有统治性优势，2000-4000 元价位段市场几乎无国外品牌竞争空间，而我们在前文分析本次国补后销量提升最大的市场即这部分市场；第三，伴随华为 Mate70 系列的发布，华为在高端市场的表现持续亮眼，而 2025 年华为在折叠屏、卫星通信、多端互联等技术上预计将进一步巩固优势，有望持续强化国产品牌概念。

以智能机为代表的消费电子走向复苏，国产品牌带动国产半导体厂商增长有望。虽

然在全球范围内消电复苏略为缓慢，但国内市场表现分化，在各地补贴和新机频出的带动下 2024 年手机市场同增持续走高，我们预计在 2025 年的国补带动下，以智能机市场为代表的消费电子在国内的复苏或将提速，受益可能性最高的国产手机品牌对国产半导体市场也有望带来较为可观的提振作用，带动消电链、安卓链芯片厂商逐步走出困境。

图52 2025 年 1 月国产手机市场排名

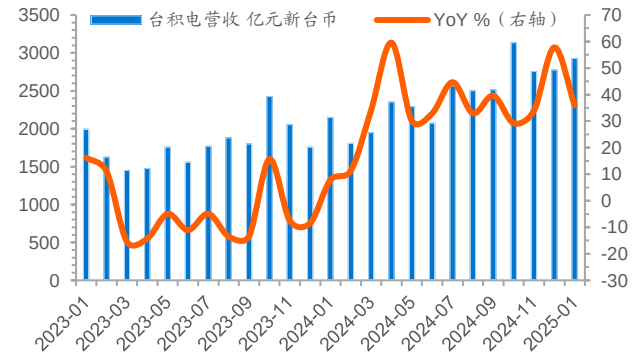


资料来源：手机中国，联储证券研究院

4.2 集成电路制造&封测：非 AI 慢慢出走底部，下游需求或拉动产业协同

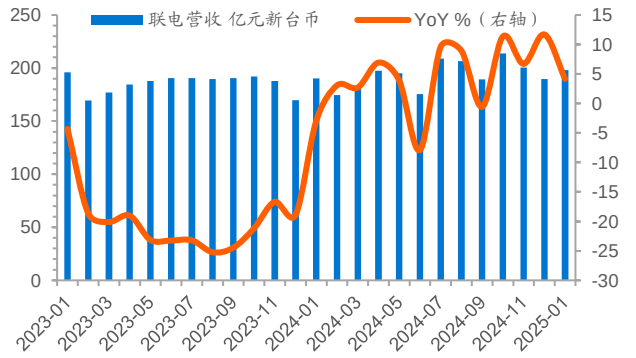
制造端来看，**各大晶圆厂在 2024 年均得到一定的复苏。**从中国台湾地区的晶圆厂月度营收视角观察，2024 年晶圆代工厂的明显变化的趋势在于：除去台积电这种大量产能和收入集中于先进制程的代工厂以外，其它如世界先进、力积电等以成熟制程为主的晶圆厂营收同比变化也已经有 5-10 个月的时间回归到增长区间段，即说明自 2024H2 以来，晶圆代工的需求是十分强劲的，且其中非 AI 的部分也回到了平稳增长的区间，我们认为在需求端大概率不会走弱的前提下，这个趋势在 2025 年是可以延续的。

图53 台积电月度营收及同比变化



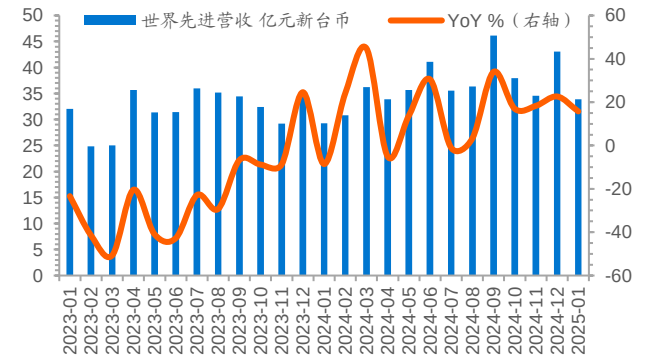
资料来源：同花顺 iFinD，联储证券研究院

图54 联电月度营收及同比变化



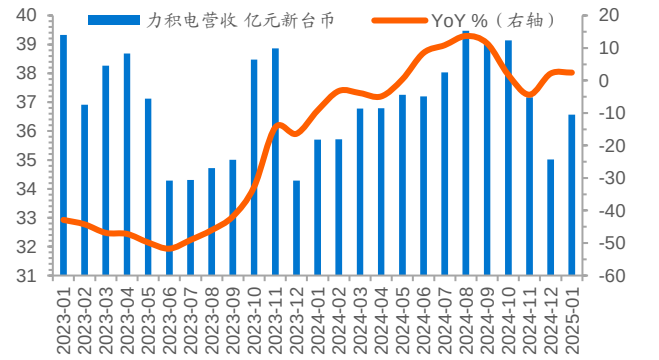
资料来源：同花顺 iFinD，联储证券研究院

图55 世界先进月度营收及同比变化



资料来源：同花顺 iFinD，联储证券研究院

图56 力积电月度营收及同比变化

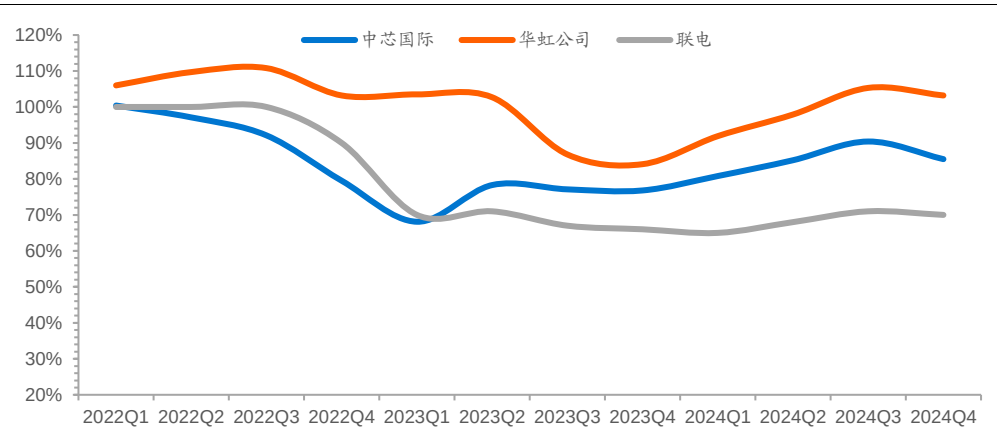


资料来源：同花顺 iFinD，联储证券研究院

国产晶圆厂的产能供给处于不断扩张中。据 TrendForce 预测，2024 年中国大陆在全球成熟节点制造产能中的占比为 34%，中国台湾为 43%。到 2027 年，中国大陆的份额预计将超过中国台湾，而韩国和美国的份额则将分别降至个位数并预计下降。据 SEMI 预测，2023 年至 2025 年间投入生产的 97 家新建制造厂中，有 57 家位于中国大陆。从具体项目来看，中芯国际深圳 12 英寸晶圆厂、华润微（润鹏）12 英寸晶圆厂、增芯 12 英寸晶圆厂等都是 2024 年新增的晶圆厂项目，此外，还有鼎泰匠芯、鹏芯微、鹏新旭等晶圆厂也在 2024 年进行了建设或投产。

同时，国产晶圆厂呈现出来的需求增长态势更加明显。我们认为相较于其它地区（尤其是中国台湾地区）的晶圆厂而言，受供应链安全的需求，大陆晶圆厂受益于国内终端需求提升的可能性更高，尤其是随着非 AI 领域的需求复苏，成熟制程和特色工艺的机会可能会更多地显现出来。我们判断在供需双端走强的情况下，国产产能的扩张与需求转向可以较好地匹配，国产晶圆厂具有较大的营收增长空间。

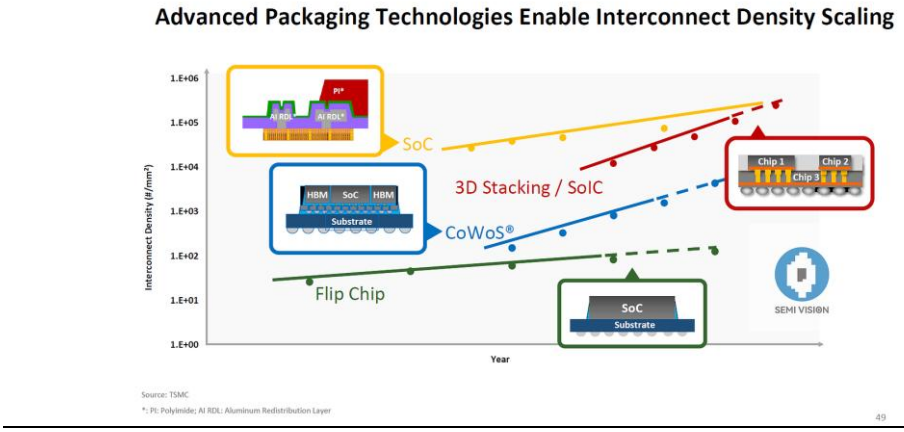
图57 主要晶圆厂产能利用率



资料来源：中芯国际财报，华虹公司财报，联电财报，联储证券研究院

封测端来看，**先进封装在提升芯片性能的意义上愈加重要。**当前集成电路的发展至摩尔定律逐渐受限的阶段，芯片性能受“存储墙”“面积墙”“功耗墙”和“功能墙”制约，摩尔定律与其说是半导体行业的技术定律，不如说是商业定律，通过缩短制程来提升芯片性能的方案技术难度大的同时且成本十分昂贵，在性能提升和成本上涨无法匹配的情况下，采取先进封装提升 IO 数量提升互联带宽达到最终芯片性能需求的方案性价比就更加凸显出来。

图58 先进封装大幅提升互联密度

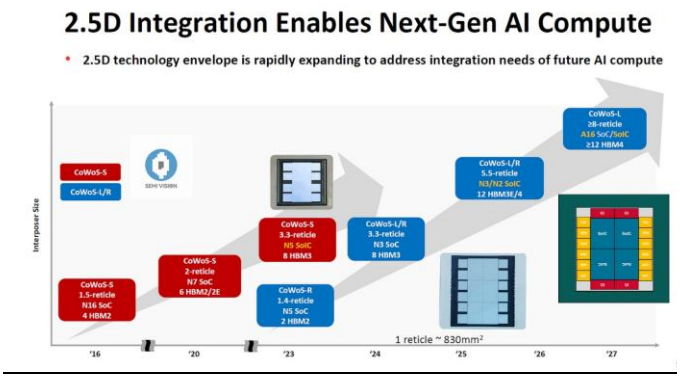


资料来源：台积电，SemiWiki，联储证券研究院

以台积电 CoWoS 为代表，先进封装的需求预计持续旺盛。在高密度异构集成中，追求高带宽和低信号延迟的互连变得越来越关键，台积电 CoWoS（Chip-on-Wafer-on-Substrate）技术因其大集成面积、高带宽内存而备受关注。

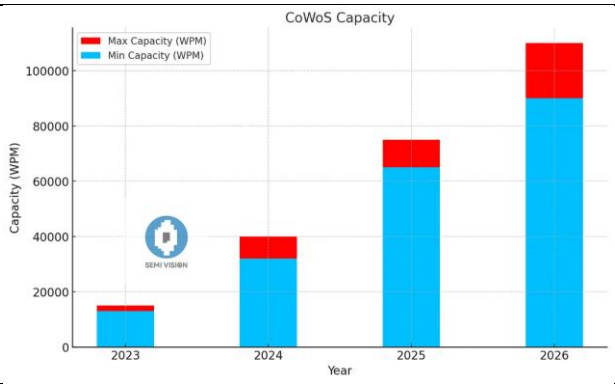
从技术演进角度，CoWoS 的演进路径清晰。由于 AI 加速器对计算性能的需求持续增长，封装中需集成更多芯片，因此中介层尺寸也随之增大，2023 年，台积电的 CoWoS 封装中介层尺寸约为 80×80mm，相当于光掩膜（Reticle）的 3.3 倍大小，其计划到 2026 年将其扩展至 100×100mm，光掩膜的 5.5 倍，并在 2027 年进一步扩大至 120×120mm，光掩膜的 8 倍。**从产能扩张角度，台积电持续扩张产能以应对 AI 发展浪潮。**2023 年台积电 CoWoS 产能约为 13000~16000WPM，预计 2025 年将达到 65000~75000WPM，2026 年预计将达到 90000~110000WPM。

图59 台积电 CoWoS 的演进路径



资料来源：台积电，SemiWiki，联储证券研究院

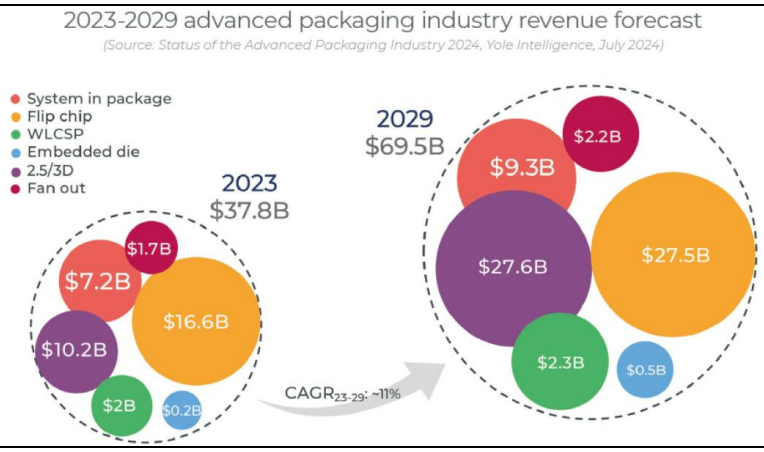
图60 台积电 CoWoS 的产能规划



资料来源：台积电，SemiWiki，联储证券研究院

对于国产 OSAT 而言，大力发展先进封装的意义显著。原因在于：①先进封装中的 2.5D、3D stacking、晶圆级封装等是 HBM、AI 芯片的重要前置技术，对于我们构造国产 AI 生态是无法绕开的；②如 chiplet 技术是整合国产不同半导体厂商产品的有效方式，有望推动国产厂商形成合力，构建产业生态；③当前在部分核心技术与设备受限严重的情况下，突破先进制程的性价比较低，因此大力发展先进封装的意义进一步提升；④DeepSeek 的落地加速国产 AI 芯片市场扩张，而其模型优化依赖高密度芯片集成，国内 AI 芯片市场占比或将提升，倒逼国产先进封装技术突破；⑤国产 OSAT 市场份额较大，技术追赶迅速，2023 年国内先进封装产能突破 150 万 WPM，长电科技和通富微电市占率合计达 37%。**预计先进封装市场将会迅速增长，前景广阔。**据 Yole 预测，全球先进封装行业的市场规模在 2023 年达到 378 亿美元，预计到 2029 年将达到 695 亿美元，2023-2029 年的 CAGR 达 10.68%。

图61 先进封装市场规模

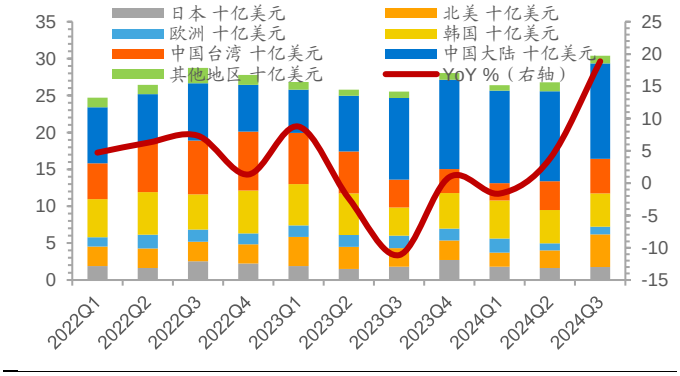


资料来源：Yole，联储证券研究院

4.3 半导体设备&材料：制造封测端的机会传导，有望实现外退内进

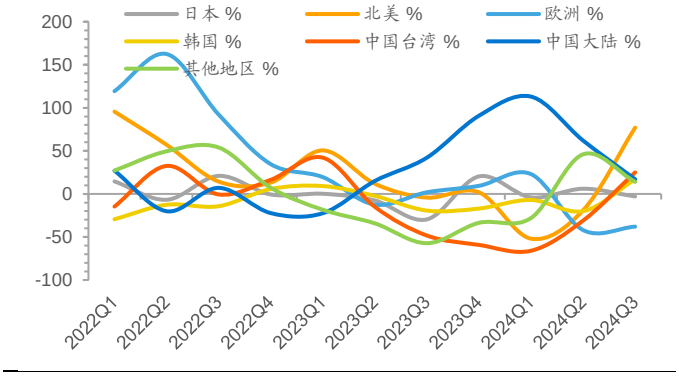
2024 年半导体设备销售一扫颓势，维持增长态势。整体看全球半导体设备销售额同比变化，进入到 2024Q2 以来全球设备销售额 YoY 回正，Q3 全球销售额达 303.7 亿美元，YoY 达 18.86%，录得近两年最高值。分结构从各地区半导体设备销售额同比变化来看，有几个特征：一是大陆地区的增速自 2023 年以来保持正值，说明了需求的可延续性；二是美国和中国台湾地区的增速在 24Q3 回正，这是由于美国推动的制造业回流和中国台湾地区受 AI 需求不断扩张带来的结果。

图62 各地区半导体设备销售额



资料来源：同花顺 iFinD，联储证券研究院

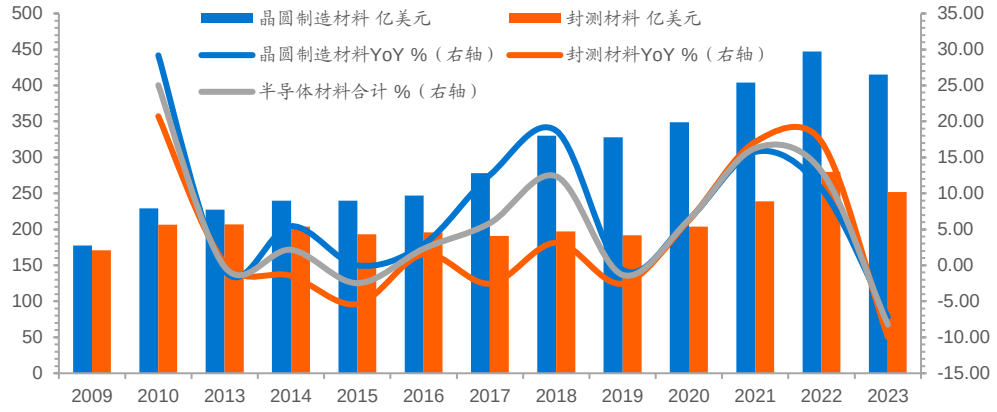
图63 各地区半导体设备销售额同比变化



资料来源：同花顺 iFinD，联储证券研究院

由于前期的低基数，半导体材料市场 2024 年有望回暖。在 2023 年由于受行业周期性调整影响，市场规模同比下降，据 SEMI 数据显示，2023 年全球半导体材料销售额为 667 亿美元，较 2022 年 726.9 亿美元下降 8.25%。受益于 AI 驱动和晶圆厂扩产等因素，预计 2024 年全球半导体材料市场将实现回暖。

图64 半导体材料销售额及同比变化



资料来源：SEMI，同花顺 iFinD，联储证券研究院

从晶圆厂 CapEx 看设备材料边际的需求变化，国产材料设备预计将维持增长。由于大陆晶圆厂在当前全球格局下是更受国产 IC 设计厂商“Local for Local”青睐的，因此其对于产能扩张的需求是高于其它地区的晶圆厂的，中芯国际和华虹公司的产能利用率持续高位运行可以予以佐证。在当前时点下，我们认为这一关键趋势未发生改变，对国产材料设备商而言，从中芯国际和华虹公司的 CapEx 动向来看，中芯国际预计资本开支将大致持平，华虹则是由于产能利用率始终保持 100%左右的高位，其新产线产能扩张的需求较大，因此我们判断 2025 年国内晶圆厂 CapEx 有望实现稳中有增。

在需求端整体稳定的情况下，我们认为结构性的调整是设备材料增长的最关键动能。从当前设备和材料的国内自给率来看，我们看到两个特点：①当前的材料设备国产化率较此前已有明显提升，根据 TechInsight，2018 年中国设备市场的自给率为 15.9%，到了 2023 年增至 23.3%，半导体材料中 2022 年硅片国产化率仅为 9%，至 2024 年 8 英寸硅片国产化率达 55%；②尽管有所提升，但当前值整体仍然偏低，因此我们认为设备材料的市场空间是依旧巨大的，TechInsight 预测 2027 年设备自给率将上升至 26.6%，当前的全球竞争环境对国产厂商而言是风险更是机遇，我们认为自主可控长期的趋势是不会发生改变的，以国产 AI 为代表的下游需求领域正处于高速发展的快车道上，有望推动设备材料厂商持续切入。

表9 主要半导体设备自给情况

设备类型	自给率	国产厂商	海外厂商
去胶	75-90%（中低端） <30%（高端）	屹唐半导体、北方华创、盛美上海、浙江宇谦、上海稷以等	日立高新技术（日）、拉姆研究（美）等
清洗	50-60%	盛美上海、北方华创、至纯科技、芯源微、屹唐半导体等	迪恩士（日）、东京电子（日）、拉姆研究（美）等
蚀刻	50-60%（成熟制程） <15%（先进制程）	中微公司、北方华创、嘉芯半导体、屹唐半导体、拓荆科技、盛美上海、芯源微等	应用材料（美）、拉姆研究（美）、东京电子（日）等
热处理	30-40%	北方华创、晶盛机电、中微公司、拓荆科技、嘉芯半导体等	ASML（荷兰）、应用材料（美）、拉姆研究（美）、东京电子（日）等
PVD	15-20%（成熟制程） 约 10%（先进制程）	北方华创、捷佳伟创、嘉芯半导体、中电科、科瑞设备有限公司、中科院沈阳科学仪器、合肥科晶材料等	ASML（荷兰）、应用材料（美）、拉姆研究（美）、东京电子（日）等
CVD/ALD	5-10%	北方华创、晶盛机电、中微公司、盛美上海、拓荆科技、嘉芯半导体等	ASML（荷兰）、应用材料（美）、拉姆研究（美）、东京电子（日）等
CMP	15-25%（成熟制程） 小于 10%（先进制程）	盛美上海、华海清科、中国电科、鼎龙控股、晶亦精微等	杜邦（美）、Thomas West（美）、JSR（日）等
涂胶显影	10-15%（成熟制程） 10%（先进制程）	盛美上海、芯源微、北方华创、中微公司、华峰测控等	陶氏化学（美）、JSR（日）、TOK（美）等
离子注入	10-20%（成熟制程） <5%（先进制程）	凯世通、中国电科、烁科中科信、北方华创、中微公司等	应用材料（美）、亚舍利（美）等
曝光	10-15%（成熟制程） 0-1%（先进制程）	上海微电子、中国电科、北方华创等	ASML（荷兰）、佳能（日）、尼康（日）等
量测	10-15%（成熟制程） <5%（先进制程）	上海微电子、中科飞测、精测电子、华海清科、北方华创等	KLA（美）、Santec Holdings（日）等

资料来源：TrendForce，联储证券研究院

表10 主要半导体材料自给情况

材料类型	自给率	国产厂商	海外厂商
硅材料	55%（8 英寸） 10%（12 英寸）	沪硅产业、中环股份、立昂微、中晶科技等	信越化学（日）、迪恩士（日）等
工艺化学品	约 10%（G3 及以上）	江化微、格林达等	霍尼韦尔（美）、住友化学（日）等
光掩膜	晶圆厂自产为主	清溢光电、路维光电、菲利华等	各大晶圆厂、Toppan（日）等
光刻胶	约 10%（高端）	华懋科技、彤程新材、南大光电、晶瑞电材、上海新阳等	JSR（日）、东京电子（日）等
CMP 抛光材料	约 30%（抛光液） 约 20%（抛光垫）	鼎龙股份、安集科技等	陶氏化学（美）、卡博特（美）等
电子气体	约 15%	华特气体、金宏气体、雅克科技等	空气化工（美）、林德集团（德国）等
靶材	自给率较高	江丰电子等	霍尼韦尔（美）、日矿金属（日）等
引线框架	约 40%	康强电子、博威合金等	住友集团（日）、三井化学（日）等
封装基板	约 20%	深南电路、兴森科技、欣兴电子等	揖斐电（日）、三星电机（韩）等
环氧塑封料	约 30%	华海诚科、联瑞新材等	住友电木（日）、日东电工（日）、日立化工（日）等
键合丝	约 30%	一诺电子、康强电子等	田中电子（日）等

资料来源：爱集微，北京半导体行业协会，联储证券研究院

5. 投资建议

进入 2025 年以来，半导体行业利好频出，我们认为 2025 年在多重因素共振下行

业有望实现进一步增长，迎来复苏的新阶段。在当前趋势下，从寻找各环节的最大公约数出发，我们看好以下方向：

云端 AI：算力需求增长依旧强劲，科技进步主体变化带来的边际影响或较为显著。预训练的算力膨胀方兴未艾，后训练和推理侧算力重要性愈发明显，成为维持行业快速增长斜率的关键动能，DeepSeek 的横空出世促使国产软硬件进一步结合，国产云端 AI 生态开辟新场景，有望实现对海外算力、模型、系统之间闭环关系的解耦，构建国产 AI 产业链的良性循环体系，建议关注：国产算力产业链。

端侧 AI：边缘 AI 设备的普及与端侧算力需求释放，或催生新的增长点。下一阶段的三大重点即：手机与 PC 市场的回暖，搭载端侧 AI 实现应用场景落地；AIoT 产品不断创新，以延伸人体关键感觉器官实现功能解放为目标；汽车领域的智能化普及趋势加速，全民智驾初步走入现实。整体来看端侧大模型发挥空间十足，看好具有增长确定性的细分赛道，建议关注：SoC、MCU、电源管理、智驾芯片、CIS 等。

自主可控：地缘政治与供应链重构背景下，中国半导体产业将加速技术突破与产能布局，同时全球产业链多元化趋势也将重塑行业竞争格局。消电国补如约而至，非 AI 领域有望逐渐走出底部区间，设备材料国产替代趋势未改，制造封测“Loca for Local”有望受益，建议关注底部反转的可能性以及自主可控进程较快的子行业，包括：半导体设备、晶圆代工、先进封装等。

6. 风险提示

下游需求复苏不及预期，若手机、PC 等需求不可延续，车用、工控未得到改善，需求整体弱化影响下，半导体全产业链都将难以实现增长；

国产 AI 发展速度不及预期，若国内 AI 企业的研发停滞，软件端应用场景落地困难，硬件端协同难度将大大提升；

地缘政治风险加剧，若上游材料设备进口难度加大，下游商品海外需求减弱，或会导致供应链风险提升。

免责声明

联储证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，但本公司及其研究人员对该等信息的准确性及完整性不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，可能会随时调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议,本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的任何观点均精准地反映了我们对标的证券和发行人的个人看法，结论不受任何第三方的授意或影响。我们所得报酬的任何部分无论是在过去、现在及将来均不会与本报告中的具体投资建议或观点有直接或间接联系。

投资评级说明

投资建议的评级标准		评级	说明
评级标准为报告发布日后的 6 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准。	股票评级	买入	相对同期基准指数涨幅在 10%以上
		增持	相对同期基准指数涨幅在 5%~10%之间
		中性	相对同期基准指数涨幅在-5%~+5%之间
		减持	相对同期基准指数跌幅在 5%以上
	行业评级	看好	相对表现优于市场
		中性	相对表现与市场持平
		看淡	相对表现弱于市场

联储证券研究院

青岛

地址：山东省青岛市崂山区香港东路 195 号 8 号楼 11、15F

邮编：266100

北京

地址：北京市朝阳区安定路 5 号院中建财富国际中心 25F

邮编：100029

上海

地址：上海市浦东新区滨江大道 1111 弄 1 号中企国际金融中心 A 栋 12 层

邮编：200135

深圳

地址：广东省深圳市南山区沙河街道深云路 2 号侨城一号广场 28-30F

邮编：518000