

# PLAY TO GENERALIZE: LEARNING TO REASON THROUGH GAME PLAY

000  
001  
002  
003  
004  
005 **Anonymous authors**  
006 Paper under double-blind review  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1098  
1099  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1198  
1199  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1298  
1299  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1398  
1399  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1498  
1499  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1598  
1599  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1698  
1699  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1798  
1799  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1898  
1899  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908<br

054 AI agents, too, have benefited from games resembling aspects of human play. These environments  
 055 encourage exploration, robustness to sparse rewards, and learning from multimodal inputs. For ex-  
 056 ample, emergent tool use has been observed in agents trained via hide-and-seek (Baker et al., 2019),  
 057 and Atari gameplay has been incorporated into training generalist agents (Reed et al., 2022). By  
 058 learning in these environments, AI systems develop robust and transferable reasoning capabilities.

059 Recent work has shown that post-training with Reinforcement Learning (RL) can unlock reasoning  
 060 behaviors from their base models (DeepSeek-AI, 2025; OpenAI, 2024). These RL-trained models  
 061 can “think before they speak”, generating internal chain-of-thought traces before outputting a final  
 062 answer. More importantly, growing evidence suggests that RL often generalizes more robustly to  
 063 out-of-distribution samples than supervised fine-tuning (SFT). For example, models trained with  
 064 RL on mathematics transfer their reasoning skills to physics (Meng et al., 2025), and navigation  
 065 agents adapt to novel environments beyond their training domains (Chu et al., 2025a). Motivated  
 066 by these findings, we ask: *since games already serve as a natural medium through which humans  
 067 acquire reasoning strategies, can post-training multimodal LLMs on gameplay similarly enhances  
 068 their ability to reason across diverse tasks?*

069 The results are striking (Fig. 1). We show that post-training a 7B-parameter multimodal LLM,  
 070 Qwen2.5-VL-7B (Bai et al., 2023), to play simple arcade-style games like Snake (Kamradt, 2025)  
 071 yields two surprising outcomes: (1) the model generalizes to previously unseen games (Sec. 2.3);  
 072 and (2) it exhibits strong reasoning abilities on multimodal math benchmarks like MathVista (Lu  
 073 et al., 2024), and multi-domain QA like MMMU (Yue et al., 2024a). Despite never observing  
 074 worked solutions, equations, or diagrams during RL post-training, the model achieves competitive  
 075 results not only against large-scale industrial systems like GPT-4o (Hurst et al., 2024), but also  
 076 against specialist models post-trained on math datasets (Tabs. 2 and 3). Furthermore, it improves  
 077 on reasoning benchmarks without degrading general visual understanding, a common limitation of  
 078 domain-specialist training (Tab. 9). Overall, gameplay emerges as an effective surrogate task for  
 079 incentivizing reasoning in multimodal LLMs.

080 Why does it work? Our ablation studies suggest that reasoning skills incentivized by gameplay can  
 081 be helpful to other multimodal reasoning tasks. For example, Snake, a game set on a *2D grid* where  
 082 the player maneuvers the “snake” to avoid collisions and collect apples, significantly improves per-  
 083 formance on math problems involving *2D coordinates*. In contrast, Rotation, a puzzle requiring  
 084 recognition of *3D object rotation angles*, more strongly boosts performance on *geometry questions  
 085 involving angles and lengths* (Fig. 3). Furthermore, jointly training on both games yields consist-  
 086 ently stronger results on downstream benchmarks than training on either game alone, suggesting  
 087 the **compositionality** of the acquired skills (Tab. 2).

088 All these results point to a new post-training strategy: rather than relying solely on domain-specific  
 089 datasets, we can design scalable and controllable **surrogate tasks for post-training**, such as games,  
 090 that unlock reasoning behaviors transferable to downstream applications. Synthetic game environ-  
 091 ments offer structured, rule-based rewards and fine-grained controllability, while also scaling far  
 092 more easily than human-annotated data. This promising paradigm of post-training with surrogate  
 093 tasks reminisces self-supervised pre-training in vision and language (He et al., 2020; Doersch et al.,  
 094 2015; Radford et al., 2018), where carefully designed pretext tasks produce broad generalization.

## 095 2 REINFORCEMENT LEARNING ON VISUAL GAMES

096 We introduce ViGaL, a novel post-training paradigm designed to enhance generalization capabilities.

### 100 2.1 GAME ENVIRONMENT

101 As show in Fig. 2, under our ViGaL paradigm, the model is trained in a game environment where  
 102 it receives states from game environment, outputs next actions, and obtains rewards as feedback  
 103 from the environment. Formally, each task, given an instruction  $I$ , can be formulated as a partially  
 104 observable Markov decision process (POMDP):  $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, \Omega)$ , where  $\mathcal{S}$  is the set of possible  
 105 environment states,  $\mathcal{O}$  is the set of observations available to the model, and  $\mathcal{A}$  represents actions  
 106 model can do in this game environment.  $T : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  is the state transition function, while  
 107  $R$  is a binary reward from the environment representing the correctness of action. Due to partial  
 108 observability, the agent perceives only observations  $o = \Omega(s)$ .

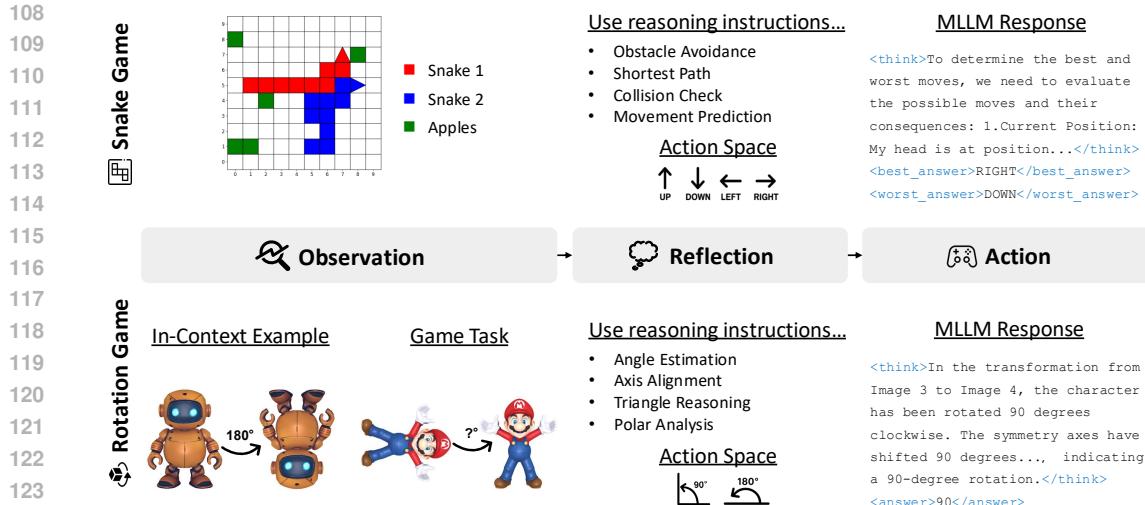


Figure 2: **Post-training MLLMs to reason through RL with games.** We propose post-training MLLMs via RL by playing visual games. We demonstrate this with two games: the classic arcade game Snake (Kamradt, 2025), and Rotation, a self-designed task to investigate spatial reasoning. In each game, the model receives multimodal inputs and follows reasoning instructions, *e.g.*, path planning in Snake, angle estimation in Rotation. It reflects to choose an action, outputs its chain-of-thoughts and decision, *e.g.*, best/worst move or predicted angle, and receives a reward. Through gameplay, the model obtains reasoning abilities that transfer to downstream multimodal reasoning tasks such as math and multi-discipline question answering.

**Snake and Rotation Games.** We design two complementary games, Snake and Rotation, to study the proposed paradigm (Fig. 2), each focusing on different MLLM capabilities. The Snake game, inspired by prior work showing that competition can enhance reasoning in MLLMs (Du et al., 2023), emphasizes strategic decision-making. We set up a dual-snake game based on SnakeBench (Kamradt, 2025), where each model independently controls one snake. The objective is to reach apples, score points, and outcompete the opponent. At time  $t$ , the environment state  $s^t$  includes the coordinates of both snakes ( $x_{s_i}^t, y_{s_i}^t$ ) for  $i \in \{1, 2\}$ , the apple location ( $x_a^t, y_a^t$ ), and the previous actions  $A_i^{t-1}$ . All elements are placed on a  $10 \times 10$  board. Each snake then selects its next action  $A_i^t \in \{\text{up, down, left, right}\}$ . A snake dies if it collides with itself, the other snake, or the board boundary; the survivor wins, or in the case of simultaneous death, the higher score decides. Unlike SnakeBench, which uses only text to represent states, we provide both images of the game board and textual descriptions as observations  $o^t = \Omega(s^t)$  for richer input. The Rotation game, inspired by rotation-angle prediction as a pre-text task in self-supervised learning (Gidaris et al., 2018), evaluates visual perception and spatial reasoning. The model is presented with two views of the same 3D object: an initial view  $I_{\text{init}}$  and a rotated view  $I_{\text{rot}}$ , obtained by rotating the object  $90^\circ$  or  $180^\circ$  around the  $z$ -axis (pointing toward the viewer). The task is to identify which rotation angle transforms  $I_{\text{init}}$  into  $I_{\text{rot}}$ . To guide reasoning, we include an in-context example with a known rotation. As in the Snake game, observations combine images and text. Together, these two games allow systematic exploration of reasoning and perception, two fundamental aspects of MLLM abilities.

## 2.2 RULE-BASED REINFORCEMENT LEARNING

We apply rule-based RL to directly post-train MLLMs for visual games, without relying on supervised learning as a warm up. The algorithm is described as follows:

**Reward design.** We use a simple rule-based reward function to avoid reward hacking (Gao et al., 2022) and help the model learn how to play the games effectively. This reward function has two components: an accuracy reward and a format reward. The total reward  $r$  is computed as the sum of an accuracy reward and a format reward  $r = r_{\text{accuracy}} + r_{\text{format}}$ . The accuracy reward  $r_{\text{accuracy}}$  is 1 if the answer is correct, and 0 otherwise. Details of reward for each game are in Appendix Sec. A.3.

Model	Wins (10)	Model	Acc. (%)	Game	ViGaL	Qwen2.5-VL-7B
ViGaL vs.		ViGaL	<b>71.9</b>	Space Invaders	280.0	85.0
Qwen2.5-VL-7B	9	Qwen2.5-VL-7B	47.4	Ms. Pacman	1370.0	670.0
Qwen2.5-VL-72B	7	Qwen2.5-VL-72B	52.1	Seaquest	80.0	60.0
Llama-4-Maverick	7	Llama-4-Maverick	66.2	Alien	540.0	450.0
Gemini-2.5-Pro	8	Gemini-2.5-Pro	51.0	Frogger	7.0	5.0
Claude-3.7-Sonnet	6	Claude-3.7-Sonnet	65.6	Breakout	0.0	9.0
GPT-4o	8	GPT-4o	61.5	Pong	-26.0	-26.0
o4-mini	6	o4-mini	70.8	Cumulative Reward	<b>2251.0</b>	1253.0

(a) Snake game.

(b) Rotation game.

(c) Atari game.

Table 1: **Game Performance.** (a) ViGaL gets high win rates (6-9 wins out of 10 matches) on Snake playing against advanced proprietary models. (b) ViGaL shows best performance on Rotation. (c) ViGaL trained on Snake and Rotation shows zero-shot generalization to unseen Atari games, achieving a nearly *doubled* cumulative reward compared to its base model (Qwen2.5-VL-7B).

**Advantage estimation and policy update.** We employ REINFORCE Leave-One-Out (RLOO) algorithm (Kool et al., 2019; Ahmadian et al., 2024) in our RL training phase. Following Group Policy Gradient (Chu et al., 2025b), we omit KL divergence regularization. Without KL constraints limiting policy changes, the model explores the solution space more freely, potentially discovering better reasoning strategies. This enables more flexible adaptation during RL training.

### 2.3 IMPLEMENTATION AND EVALUATION ON GAMES

**Implementation details.** We employ Qwen2.5-VL-7B-Instruct (Bai et al., 2023) as our base model. We follow DeepSeek-R1 (DeepSeek-AI, 2025), using a combination of rule-based format rewards and accuracy rewards, with RLOO (Kool et al., 2019; Ahmadian et al., 2024) as the core RL algorithm. We implement our training within a multimodal input RL framework based on Open-RLHF (Hu et al., 2024). For hyperparameters, we adopt the default settings from MM-Eureka (Meng et al., 2025), including a global batch size of 128, a rollout batch size of 128, a rollout temperature of 1.0, and a learning rate of  $1e^{-6}$ . Training uses 6 A100-80G GPUs.

**Game training data.** We build game environments to collect training data for our experiments. For Snake, we leverage SnakeBench (Kamradt, 2025) as our data engine. For Rotation, we utilize Hunyuan3D (Team, 2025), which generates 3D meshes from images or text instructions. We render each mesh into 2D images from different orientations, creating image pairs with associated rotation angles as ground truth labels for RL training. Our comprehensive data generation pipeline enables producing training samples at any desired scale with fully customized settings. For experiments, we synthesize 36K samples per game, sufficient for convergence. Details are in Appendix Sec. A.1.

**Competing with leading models on Snake and Rotation.** To evaluate the game capabilities of ViGaL models, we initialize environments in diverse states unseen during training. For Snake (Tab. 1a), we randomly initialize games 10 times with two models competing directly, measuring win counts. For Rotation (Tab. 1b), we measure rotation angle prediction accuracy on comprehensive validation sets with 3D object meshes unseen during training. Our 7B-parameter model consistently outperforms proprietary models in both games. Results confirm that RL effectively unlocks small 7B models’ ability to excel in visual games requiring environmental understanding, reasoning, planning, and interactive decision-making.

**Out-of-distribution generalization to Atari games.** We then test ViGaL on Atari-GPT (Waytowich et al., 2024), a benchmark for evaluating MLLMs as decision-making agents in Atari video games such as in Fig. 5. The benchmark consists of seven different Atari games, with detailed settings in Appendix Sec. B.1. We follow most settings and prompts from Atari-GPT, with a small modification to ensure format correctness for all models. Following Atari-GPT (Waytowich et al., 2024), we report cumulative reward over 1K steps as the evaluation metric, where higher rewards indicate better performance. As shown in Tab. 1c, ViGaL shows significant cumulative reward improvement on Atari games despite being trained only on Snake and Rotation games. This is particularly notable because Atari games differ substantially from our training games in both visual appearance and gameplay strategies. These results suggest that our rule-based RL training approach enables strong generalization to previously unseen game environments.

Model	Math					Geometry		
	Avg.	Avg.	MathVista	MathVerse	MathVision	Avg.	GeoMath	Geo3K
Proprietary Model								
GPT-4o (Hurst et al., 2024)	47.5	47.3	61.4	50.2	30.4	46.8	50.2	43.5
Gemini-2.0-Flash (Team, 2023)	55.4	56.4	73.4	54.6	41.3	54.4	55.3	53.5
Multimodal Reasoning Model Post-Trained on Qwen2.5-VL-7B (Bai et al., 2023)								
<i>Base Model (Qwen2.5-VL-7B)</i>	46.3	47.7	68.0	49.0	26.0	44.8	44.0	45.6
R1-Onevision-7B (Yang et al., 2025)	40.9	46.8	64.1	46.4	<b>29.9</b>	35.0	45.4	24.5
R1-VL-7B (Chen et al., 2025b)	40.9	42.7	63.5	40.0	24.7	39.0	42.0	36.1
MM-Eureka-Qwen-7B (Meng et al., 2025)	39.3	50.1	73.0	50.3	26.9	28.4	53.1	3.8
Reason-RFT-Zero-7B (Tan et al., 2025)	46.5	38.1	60.7	35.3	18.3	54.9	55.0	54.8
VLAA-Thinker-7B (Chen et al., 2025a)	51.3	48.7	68.0	51.7	26.4	53.9	51.1	56.6
OpenVLThinker-7B (Deng et al., 2025)	52.1	47.8	70.2	47.9	25.3	56.4	49.2	63.5
ViGaL Snake	51.6	49.4	70.7	51.1	26.5	55.0	49.9	60.0
ViGaL Rotation	52.8	49.3	71.2	50.4	26.3	<b>57.9</b>	51.7	64.1
ViGaL Snake + Rotation	<b>53.9</b>	<b>50.6</b>	71.9	52.4	27.5	57.1	51.0	63.3
	±0.3	±0.3	±0.4	±0.2	±0.3	±0.5	±0.3	±0.4

Table 2: **Results on multimodal mathematical benchmarks.** We compare to other multimodal reasoning models. Results post-trained on the same subject as the evaluation are de-emphasized, while our ViGaL models only use games for post-training. **Bold** numbers are the best in each Avg. column. We include standard deviations of three independent runs for ViGaL Snake + Rotation.

Model	CLEVR <sup>+</sup>				Multi-Discipline		
	Avg.	Avg.	CLEVR-M	S-CLEVR	Avg.	MMMU <sub>val</sub>	MMMU-Pro <sub>overall</sub>
Proprietary Model							
GPT-4o (Hurst et al., 2024)	55.9	51.2	68.1	34.3	60.5	69.1	51.9
Gemini-2.0-Flash (Team, 2023)	–	46.3	64.9	27.6	–	71.9	–
Multimodal Reasoning Model Post-Trained on Qwen2.5-VL-7B (Bai et al., 2023)							
<i>Base Model: Qwen2.5-VL-7B</i>	50.3	54.9	74.6	35.2	45.7	54.3	37.0
R1-Onevision-7B (Yang et al., 2025)	53.7	65.1	75.5	54.7	42.3	51.9	32.6
R1-VL-7B (Chen et al., 2025b)	53.9	68.0	87.4	48.6	39.7	50.0	29.4
MM-Eureka-Qwen-7B (Meng et al., 2025)	62.8	79.3	<b>98.4</b>	60.1	46.4	55.8	36.9
Reason-RFT-Zero-7B (Tan et al., 2025)	58.6	76.2	99.4	53.0	40.9	51.2	30.6
VLAA-Thinker-7B (Chen et al., 2025a)	61.7	<b>83.4</b>	94.7	72.1	40.1	48.2	31.9
OpenVLThinker-7B (Deng et al., 2025)	60.4	82.4	93.8	71.0	38.5	54.8	22.1
ViGaL Snake	64.4	82.6	92.6	72.6	46.2	55.8	36.6
ViGaL Rotation	63.3	80.7	93.0	68.3	45.9	54.1	37.7
ViGaL Snake + Rotation	<b>64.7</b>	81.7	91.9	71.4	<b>47.7</b>	58.0	37.4

Table 3: **Results on multimodal spatial and multi-discipline reasoning benchmarks.** CLEVR-M denotes CLEVR-Math (Lindström & Abraham, 2022), and S-CLEVR stands for Super-CLEVR (Li et al., 2023). Results post-trained on the same subject as the evaluation are de-emphasized, while ViGaL is exclusively post-trained using games. **Bold** numbers are the best in each Avg. column.

### 3 VISUAL REASONING GENERALIZATION

**Evaluation collection.** Following prior studies (Tong et al., 2024a; Li et al., 2024c), we systematically divide existing benchmarks into two broad categories: (i) *reasoning-oriented benchmarks* requiring multi-step or mathematical reasoning, and (ii) *general-purpose perception benchmarks* assessing visual understanding and perception abilities.

For reasoning-oriented evaluation, we test on four key areas: *Math* (MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024b)), *Geometry* (GeoMath (Gao et al., 2023; Shi et al., 2024), Geometry3K (Lu et al., 2021)), *CLEVR+* (CLEVR-Math (Lindström & Abraham, 2022), Super-CLEVR (Li et al., 2023)), and *Multi-Discipline* (MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b)). For general perception, we evaluate across three cate-

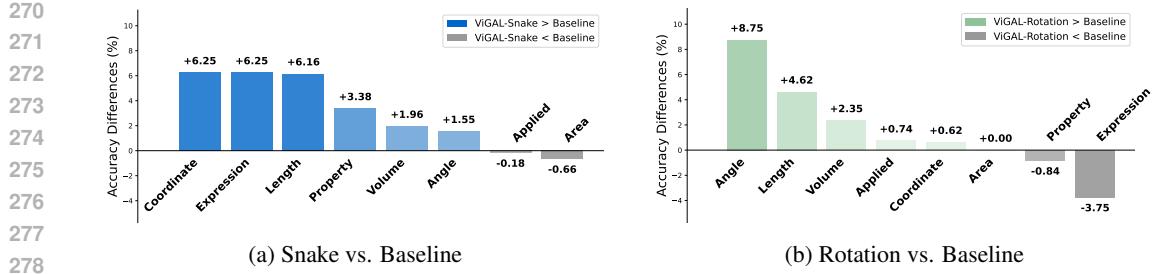


Figure 3: **Per-category gains on MathVerse are *not* uniform.** The eight math categories follow MathVerse (Zhang et al., 2024). (a) Snake yields the largest gains on *Coordinates* and *Expressions*, consistent with its 2D grid structure. (b) Rotation boosts *Angle* and *Length* questions but reduces *Expression* accuracy, suggesting its training primarily incentivizes orientation recognition.

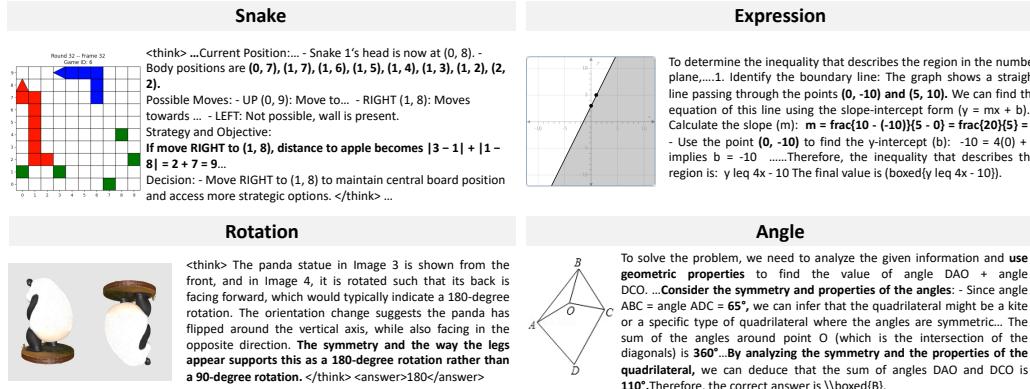


Figure 4: **Reasoning trace of different games and math questions.** Top: Algebraic functions and coordinate-level interpretations that emerge from playing the Snake game help solving *Expression* questions. Bottom: Spatial reasoning skills incentivized by playing the Rotation game appear when solving *Angle*-related problems.

gories: General (MuirBench (Wang et al., 2024a), CRPE (Kazemzadeh et al., 2014)), Vision-Centric (MMVP (Tong et al., 2024b), RealWorldQA (X.AI, 2024), MMStar (Chen et al., 2024a), MME (Fu et al., 2023), BLINK (Fu et al., 2024)), and OCR & Chart (AI2D (Kembhavi et al., 2016), SEED-Bench-2-Plus (Li et al., 2024a), DocVQA (Mathew et al., 2021), OCRBench (Liu et al., 2023)). More detailed descriptions of each benchmark are provided in Appendix B.6.

### 3.1 MAIN RESULTS

**Zero-shot generalization from gameplay to multimodal reasoning.** Our approach consistently shows remarkable generalization capabilities on mathematical and other reasoning tasks, despite having no direct exposure to in-domain training data during RL post-training. As shown in Tab. 2, our method notably outperforms models specifically RL-trained on mathematical tasks. For instance, ViGaL Snake + Rotation achieves 0.5% higher accuracy than MM-Eureka-Qwen-7B (Meng et al., 2025) 28.7% on Geometry, even though MM-Eureka-Qwen-7B was explicitly trained on high-quality mathematical and geometry datasets.

This strong generalization extends beyond mathematics. Tab. 3 shows that ViGaL Snake + Rotation outperforms R1-OneVision-7B (Yang et al., 2025) by 5.4% on average across MMMU series benchmarks, which test multi-disciplinary reasoning. This is particularly notable since R1-OneVision-7B was trained on a carefully curated comprehensive dataset spanning multiple subjects.

These empirical results suggest that gameplay-based post-training develops fundamental reasoning capabilities that transfer more effectively than direct RL training on diverse task-specific datasets. Moreover, the gameplay environment appears to encourage general problem-solving strategies that consistently generalize well to out-of-domain tasks.

Model	Avg.	MathVista	MathVerse	MathVision
Base Model: Qwen2.5-VL-7B	47.7	68.0	49.0	26.0
MM-Eureka-Qwen-7B	50.1	73.0	50.3	26.9
ViGaL Snake + Rotation	50.6	71.9	52.4	27.5
ViGaL Snake + Rotation + Math Data	51.8	72.3	54.5	27.7

Table 4: **Gameplay complements math data.** Adding math data MMK12 on top of ViGaL yields further gains in math performance. With access to the same amount of math data, ViGaL outperforms MM-Eureka (Meng et al., 2025) on average of the three math benchmarks.

**Blending multiple games enhances generalization.** As shown in Tab. 2, post-training on Snake achieves best performance on the CLEVR+ benchmark, while training on Rotation yields stronger results on geometry reasoning. Furthermore, training on both Snake and Rotation enables learning complementary skills, improving the overall benchmark average to 63.1%. These findings indicate that combining game environments drives meaningful performance gains, demonstrating Visual Gaming Learning’s potential as a promising paradigm for enhancing generalizable reasoning without large-scale domain-specific data. Expanding the types of games during training consistently scales performance across different visual reasoning tasks.

**Different games benefit distinct math subfields.** To study which types of problems in the math benchmarks benefit from game play, we analyze accuracy differences across MathVerse (Zhang et al., 2024) subcategories between ViGaL models trained with Snake or Rotation, as shown in Fig. 3. We find that training on the Snake game significantly improves performance on the sub-categories like Expressions and Coordinates, while training on Rotation notably enhances performance on questions about angles and lengths. To understand why different games help with different types of math, we compare the reasoning processes required for playing games versus solving math problems. As shown in Fig. 4, solving Expressions questions involves algebraic functions and coordinate-level interpretations of graphical representations, which closely align with the spatial reasoning process in Snake. Similarly, solving angle-related questions is consistent with requirement of playing Rotation game to reason about rotational angles of 3D objects. These results suggest that playing different games develops fundamental skills like spatial modeling and algebraic calculation that transfer to visual math questions. To validate whether this pattern generalizes beyond Snake and Rotation, we extend this analysis to four additional games (Maze, Tetris, Sudoku, Sokoban) and identify systematic transfer patterns based on cognitive alignment between game mechanics and mathematical skills. A complete multi-game study using K-Means clustering analysis on MathVerse subject-level categories is presented in Appendix Sec. B.4. The experiment on quantitatively analyzing the correlation between math and game is in Sec. B.9 in the Appendix. Furthermore, joint training on both games leads to improvements across *all* reasoning categories (see Appendix Sec. B.3). We also include qualitative analyses on improvements in math reasoning after RL in Appendix Sec. C.

**Gameplay complements math data.** We explore the complementary benefits of adding math data to the gameplay training pipeline, for which we implement a two-stage training process. Stage 1 equals to ViGaL setup, training the model on Snake and Rotation games. In stage 2, we further finetune the stage 1 model on MMK12 (Meng et al., 2025), a multimodal mathematical reasoning dataset containing approximately 12k examples. Stage 2 training uses the identical data and settings as MM-Eureka-Qwen-7B (Meng et al., 2025). As shown in Tab. 4, the integration of mathematical data in stage 2 yields a continuous improvement of 1.2% on average across three mathematical benchmarks. This demonstrates the complementary relationship between our visual game learning approach and mathematical data post-training. Moreover, ViGaL with math data significantly outperforms MM-Eureka-Qwen-7B by 1.7% on mathematical benchmarks on average, using the same math data. These results suggest that visual game learning can serve as an effective surrogate task together with domain-specific data to improve performance on target tasks.

**Preserving general visual capabilities while reasoning enhancement.** To examine whether generalization on reasoning tasks leads to degradation in general visual capabilities, we evaluate ViGaL Snake + Rotation on a broader set of MLLM benchmarks. As shown in Tab. 9 in Appendix, compared to Qwen2.5-VL-7B prior to RL tuning, our model maintains comparable general visual performance while achieving stronger math reasoning results. In contrast, other models that improve

	(a) Text prompt design.					(b) Reward design.					(c) Difficulty control.				
	prompt	Avg.	Math	CLEVR+	Geo.	reward	Avg.	Math	CLEVR+	Geo.	difficulty control	Avg.	Math	CLEVR+	Geo.
378	base model	49.1	47.7	54.9	44.8	base model	49.1	47.7	54.9	44.8	base model	49.1	47.7	54.9	44.8
379	w/o reasoning instr.	59.5	48.0	80.4	50.1	best moves	59.6	48.2	80.4	50.2	w/o difficulty control	60.6	48.8	81.4	51.8
380	w/ reasoning instr.	<b>62.3</b>	49.4	82.6	55.0	best & worst moves	<b>62.3</b>	49.4	82.6	55.0	w/ difficulty control	<b>62.3</b>	49.4	82.6	55.0
381						w/ random label	49.4	47.5	55.4	47.5					
382															
383	(d) Data scalability.					(e) Input modality.					(f) SFT vs. RL.				
384	training samples	Avg.	Math	CLEVR+	Geo.	input modality	Avg.	Math	CLEVR+	Geo.	post-training	Avg.	Math	CLEVR+	Geo.
385	base model	49.1	47.7	54.9	44.8	base model	49.1	47.7	54.9	44.8	base model	49.1	47.7	54.9	44.8
386	16K	60.1	48.9	81.2	50.3	text	59.6	48.5	80.1	50.3	SFT	47.2	38.0	71.5	32.1
387	36K	<b>62.3</b>	49.4	82.6	55.0	vision & text	<b>62.3</b>	49.4	82.6	55.0	RL	<b>62.3</b>	49.4	82.6	55.0
388															

Table 5: **Ablation study.** We ablate different aspects of ViGaL with Snake and evaluate on downstream benchmarks. The similar evaluation with Rotation is in Sec. B.2. Each benchmark consists of several subtasks (Tab. 2 and Tab. 3), and we report their averages. The base model is Qwen2.5-VL-7B, whose results are in gray. The default settings in Tab. 2 and Tab. 3 are highlighted in blue.

math performance through RL post-training often exhibit substantial drops in general visual capabilities. These results demonstrate that our gameplay-based approach enables math generalization without compromising other visual abilities.

### 3.2 ABLATION STUDY

We ablate key design choices in the Snake environment, evaluate each variant on downstream benchmarks, and report the results in Tab. 5. The corresponding ablation for the Rotation environment is provided in Appendix Sec. B.2.

**Reasoning instructions in the text prompt help.** We use reasoning instructions, such as “finding the nearest apple by calculating Manhattan distances”, in the text prompts to guide the model thinking chains. The complete text prompts are in Appendix Sec. A.2. In Tab. 5a, we demonstrate that reasoning instructions brings a significant improvement of 2.8%, from 59.5% to 62.3%, for Snake in average accuracy over the three out-of-domain benchmarks.

**Reward design of pre-text game matters for downstream tasks.** We show that reward design of RL for games plays a crucial role for the downstream tasks. As shown in Tab. 5b, we first ask the model to predict only the best next move, defined as the action that moves toward the closest apple while avoiding death. In our improved reward design, we task the model with simultaneously predicting both the best and worst next moves, where the worst move leads directly to losing the game. More importantly, it leads to improvements across all downstream tasks, bringing an average increase of 1.8%. These results suggest that proper reward design in pre-text game can improve not only gameplay capabilities but also generalization to downstream tasks.

Furthermore, inspired by several prior works that improve model performance without labeled rewards (Zhao et al., 2025) or with random labels (Shao et al., 2025), we also provide a random reward ablation, where we still ask the model to predict both best and worst moves but use random moves as the labels. We report the results in the last row in Tab. 5b. In our gameplay setting, RL with random labels reports 49.4% on average and does not provide significant gains over the base model, different from the conclusions in prior works (Shao et al., 2025). Potential explanations lie in the difference in data domains and base models, where other works applied random labels to text-only mathematical data while our work applies random labels to visual game data.

**Controlling game difficulty for better reasoning.** Gameplay for RL post-training offers unique opportunities to easily control task difficulty. We present an ablation study on difficulty control importance. We define difficulty based on snake length, where longer snakes represent higher difficulty. For controlled difficulty, we collect training data using states where snake length falls within a moderate range of 1-5. Details are in Sec. A.1. As shown in Tab. 5c, difficulty control achieves 61.4% overall accuracy compared to 60.6% without control. This suggests our game engine can easily generate appropriately difficult data, helping prevent model sub-optimization during RL training.

**RL on games shows data scalability.** Thanks to using game engine, we can generate data at any scale with high flexibility. To show data scalability on RL of visual games, we conduct experiments using 16k and 32k snake game samples, respectively. As in Tab. 5d, scaling data from 16k to

432 32k brings a performance improvement of 1.3% on average across all domains. This suggests the  
 433 potential of the proposed ViGaL paradigm to improve downstream performance by easily scaling  
 434 training data, which contrasts with the data scaling challenges of domain-specific human annotated  
 435 data, requiring extensive manual effort.  
 436

437 **Both text and vision contribute to better visual reasoning.** To isolate the contributions of text  
 438 and vision modalities, we conduct an ablation study with a text-only setting. In this setup, we rep-  
 439 resent game states—including snake positions, apple locations, and boundary constraints—using  
 440 only textual descriptions during RL training. The model trained with text-only inputs on the Snake  
 441 game demonstrates substantial improvements across all multimodal benchmarks, with average per-  
 442 formance increasing from 49.1% to 59.6%. Incorporating visual inputs yields an additional 1.8%  
 443 performance gain. These results demonstrate that multimodal RL enhances visual reasoning capa-  
 444 bilities, with complementary contributions from both text and vision modalities.  
 445

446 **RL generalizes better than SFT from games to math.** To evaluate the out-of-domain gener-  
 447 alization of ViGaL, we compare it with supervised fine-tuning (SFT) using identical visual game  
 448 data. Tab. 5f shows that SFT with Snake game data degrades the base model’s performance on  
 449 both mathematical reasoning and geometry tasks by a notable 9.7% and 12.7%, respectively. While  
 450 SFT produces modest improvements on CLEVR+, these gains are substantially smaller than those  
 451 achieved by RL. Overall, RL improves performance by 12.3%, whereas SFT decreases performance  
 452 by 1.9%. This stark contrast demonstrates that RL better preserves and extends the model’s reason-  
 453 ing capabilities to new domains.  
 454

## 4 RELATED WORK

455 **Reinforcement Learning in MLLMs.** Reinforcement Learning (RL) increasingly enhances rea-  
 456 soning in Large Language Models (LLMs) beyond Supervised Fine-Tuning (SFT). Text-only mod-  
 457 els like DeepSeek-R1 (DeepSeek-AI, 2025) demonstrate RL’s efficacy, especially with rule-based  
 458 rewards, for complex reasoning. This paradigm is now being extended to Multimodal LLMs  
 459 (MLLMs). Recent MLLM research explores RL for improved visual reasoning, drawing from  
 460 LLM successes. Various works (Peng et al., 2025; Huang et al., 2025; Chen et al., 2025b) in-  
 461 vestigate multi-stage training, trace supervision, or rule-based RL for specific visual subdomains  
 462 like geometry and counting. Others focus on different RL algorithms like Process Reward Models  
 463 (PRMs) (Luo et al., 2025; Xiang et al., 2024), often moving beyond SFT-based Chain-of-Thought  
 464 generation (Dong et al., 2024; Thawakar et al., 2025). Many efforts favor simpler rule-based re-  
 465wards (Huang et al., 2025; Zhou et al., 2025) over complex reward models prone to hacking (Eisen-  
 466stein et al., 2023). Unlike approaches training on costly, domain-specific reasoning datasets, our Vi-  
 467 GaL paradigm extends rule-based RL to simple, synthetic visual games, demonstrating these serve  
 468 as scalable, cost-effective pre-text tasks.  
 469

470 **Generalization in MLLMs.** Achieving robust generalization to novel tasks, distributions, and  
 471 domains is central to MLLM development. RL shows promise for better out-of-distribution (OOD)  
 472 generalization compared to SFT (Chen et al., 2025b; Meng et al., 2025), and developing multi-step  
 473 reasoning like CoT (Wei et al., 2022) is itself generalization. This is often pursued through training  
 474 on large, diverse instruction-following datasets (Li et al., 2024b; Liu et al., 2024; Chen et al., 2024b)  
 475 or explicitly training general reasoning capabilities (Yang et al., 2025; Huang et al., 2025). While  
 476 these methods advance OOD generalization, they typically operate within the same broad domain  
 477 of complex visual reasoning as training data. Our ViGaL paradigm investigates stronger out-of-  
 478 domain generalization, showing fundamental skills learned from simple synthetic games transfer  
 479 zero-shot to enhance performance on entirely different, complex domains like visual mathematics  
 480 and multi-discipline questions, without domain-specific data exposure.  
 481

482 **Transfer Learning and Curriculum Learning.** The idea that training on simpler tasks can im-  
 483 prove performance on more complex ones is often framed as transfer learning. In this view, a model  
 484 first learns from a source task or domain, then reuses that knowledge to improve a different target  
 485 task (He et al., 2020). Before MLLMs, this principle was studied widely in computer vision through  
 486 self-supervised pretext tasks, where models are trained without labels to solve auxiliary tasks such

486 as predicting image rotations (Gidaris et al., 2018), relative positions of patches (Doersch et al.,  
 487 2015), or colorization, and the learned representations are then transferred to downstream tasks like  
 488 detection or segmentation. These methods show that solving carefully designed pretext tasks can  
 489 produce robust visual features that support related vision problems. In reinforcement learning, cur-  
 490 riculum learning (Bengio et al., 2009) applies a related idea by ordering tasks or data from easy to  
 491 hard so that agents gradually master more difficult behaviors. RL curricula and autocurricula have  
 492 been used to grow complex behaviors from simple environments, for example in multi-agent games  
 493 and robotics (Baker et al., 2019; Narvekar et al., 2020), typically within a single task family where  
 494 later tasks share the same goal structure as earlier ones. Our ViGaL paradigm is different in two  
 495 key respects. First, unlike self-supervised pretext approaches that transfer static feature representa-  
 496 tions within the visual domain, we transfer learned reasoning policies, such as look-ahead planning,  
 497 spatial verification, and constraint satisfaction, across tasks and domains. Second, unlike RL cur-  
 498 riculum work that mainly increases difficulty within the same environment class, ViGaL transfers  
 499 policies learned in visual control games to abstract symbolic reasoning tasks in coordinate geometry  
 500 and multi-discipline question answering, without using domain-specific supervision in those target  
 501 domains.

## 5 CONCLUSION

502 We introduced Visual Game Learning (ViGaL), a novel post-training paradigm where MLLMs learn  
 503 transferable reasoning by playing simple arcade-style games. Our core finding is that RL on games  
 504 like Snake and Rotation, *without any in-domain math data*, significantly boosts MLLM performance  
 505 on mathematical and multi-discipline benchmarks, surpassing specialized models and even large  
 506 proprietary systems. Ablations confirm the importance of game design, reward structure, and that  
 507 RL outperforms SFT, while distinct games unlock different skills. We posit that games instill fun-  
 508 damental reasoning skills, suggesting a new avenue for using scalable, controllable synthetic games  
 509 as powerful pre-text tasks to unlock generalizable reasoning. This work opens doors to exploring a  
 510 broader range of game-based learning for generalizable AI.

## 513 ETHICS STATEMENT

514 This work does not involve human subjects, private or sensitive data, or personally identifiable  
 515 information. All training and evaluation data are either synthetically generated (Appendix A.1)  
 516 or come from publicly available benchmarks (Appendix B.6). Our research adheres to the ICLR  
 517 Code of Ethics, including principles of scientific integrity, fairness, and responsible stewardship.  
 518 We believe the contributions of this work advance multimodal reasoning without raising foreseeable  
 519 ethical or societal risks.

520 **Use of Large Language Models:** An LLM was used to assist with grammar refinement of text.  
 521 Further details are provided in Appendix D.

## 523 REPRODUCIBILITY STATEMENT

525 We have taken steps to ensure the reproducibility of our results. The training environments, rein-  
 526 forcement learning setup, and hyperparameters are described in Section 2. Details of synthetic data  
 527 generation are provided in Appendix A.1, training prompts in Appendix A.2, and reward functions  
 528 in Appendix A.3. Evaluation protocols for Atari-GPT and visual reasoning benchmarks are specified  
 529 in Appendix B.1 and Appendix B.6.

## 531 REFERENCES

532 Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with  
 533 reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

535 Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin,  
 536 Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learn-  
 537 ing from human feedback in llms. In *ACL*, 2024.

539 Fahad Alhasoun and Sarah Alnaghiehish. Probabilistic programming bots in intuitive physics game  
 play. In *AAAI*, 2021.

540 Kelsey Allen, Franziska Brändle, Matthew Botvinick, Judith E. Fan, Samuel J. Gershman, Alison  
 541 Gopnik, et al. Using games to understand the mind. *Nature Human Behaviour*, 2024.

542

543 Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. Rapid trial-and-error learning with  
 544 simulation supports flexible tool use and physical reasoning. *PNAS*, 2020.

545

546 Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak.  
 547 Online difficulty filtering for reasoning-oriented reinforcement learning. *arXiv:2504.03380*, 2025.

548

549 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang  
 550 Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, local-  
 551 ization, text reading, and beyond. *arXiv:2308.12966*, 2023.

552

553 Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor  
 554 Mordatch. Emergent tool use from multi-agent autocurricula. In *ICLR*, 2019.

555

556 Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In  
 557 *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

558

559 Lara Bertram. Digital learning games for mathematics and computer science education: The need for  
 560 preregistered rcts, standardized methodology, and advanced technology. *Frontiers in Psychology*,  
 561 2020.

562

563 Franziska Brändle, Kelsey R Allen, Josh Tenenbaum, and Eric Schulz. Using games to understand  
 564 intelligence. In *CogSci*, 2021.

565

566 Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, et al. SFT or RL? an early  
 567 investigation into training R1-Like reasoning large vision-language models. *arXiv:2504.11468*,  
 568 2025a.

569

570 Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-V: Reinforcing super generalization  
 571 ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025b.

572

573 Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi  
 574 Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language  
 575 models? *arXiv:2403.20330*, 2024a.

576

577 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, et al. InternVL: Scaling up  
 578 vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024b.

579

580 Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans,  
 581 et al. SFT memorizes, RL generalizes: A comparative study of foundation model post-training.  
 582 *arXiv:2501.17161*, 2025a.

583

584 Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong  
 585 reinforcement learning baseline for model reasoning. *arXiv:2504.02546*, 2025b.

586

587 DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning.  
 588 *arXiv:2501.12948*, 2025.

589

590 Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker:  
 591 An early exploration to complex vision-language reasoning via iterative self-improvement.  
 592 *arXiv:2503.17352*, 2025.

593

594 Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by  
 595 context prediction. In *ICCV*, 2015.

596

597 Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei  
 598 Liu. Insight-V: Exploring long-chain visual reasoning with multimodal large language models.  
 599 *arXiv:2411.14432*, 2024.

600

601 Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving  
 602 factuality and reasoning in language models through multi-agent debate. In *ICML*, 2023.

594 Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D’Amour, DJ Dvijotham,  
 595 et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking.  
 596 *arXiv:2312.09244*, 2023.

597

598 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei  
 599 Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multi-  
 600 modal large language models. *arXiv:2306.13394*, 2023.

601 Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A  
 602 Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but  
 603 not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

604

605 Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, et al. G-LLaVA:  
 606 Solving geometric problem with multi-modal large language model. *arXiv:2312.11370*, 2023.

607

608 Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization.  
 609 *arXiv:2210.10760*, 2022.

610

611 Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by  
 612 predicting image rotations. *arXiv:1803.07728*, 2018.

613

614 Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for  
 615 unsupervised visual representation learning. In *CVPR*, 2020.

616

617 Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. OpenRLHF: An  
 618 easy-to-use, scalable and high-performance RLHF framework. *arXiv:2405.11143*, 2024.

619

620 Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, et al. Vision-R1: In-  
 621 centivizing reasoning capability in multimodal large language models. *arXiv:2503.06749*, 2025.

622

623 Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al.  
 624 GPT-4o system card. *arXiv:2410.21276*, 2024.

625

626 Greg Kamradt. Snake Bench: Competitive snake game simulation with LLMs. <https://github.com/gkamradt/SnakeBench>, 2025.

627

628 Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to  
 629 objects in photographs of natural scenes. In *EMNLP*, 2014.

630

631 Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali  
 632 Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

633

634 Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for  
 635 free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019.

636

637 Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus:  
 638 Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv  
 639 preprint arXiv:2404.16790*, 2024a.

640

641 Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, et al. LLaVA-NeXT-Interleave:  
 642 Tackling multi-image, video, and 3d in large multimodal models. *arXiv:2407.07895*, 2024b.

643

644 Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi,  
 645 Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint  
 646 arXiv:2408.08632*, 2024c.

647

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin  
 648 Van Durme, and Alan L. Yuille. Super-CLEVR: A virtual benchmark to diagnose domain ro-  
 649 bustness in visual reasoning. In *CVPR*, 2023.

650

651 Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-Math: A dataset for compositional  
 652 language, visual and mathematical reasoning. In *IJCLR*, 2022.

648 Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee.  
 649 LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. <https://llava-vl.gitub.io/blog/2024-01-30-llava-next/>, 2024.  
 650

651 Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin,  
 652 Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large  
 653 multimodal models. *arXiv:2305.07895*, 2023.  
 654

655 Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu.  
 656 Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning.  
 657 *arXiv:2105.04165*, 2021.  
 658

659 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, et al. MathVista:  
 660 Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*,  
 661 2024.  
 662

663 Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, et al.  
 664 URSA: Understanding and verifying chain-of-thought reasoning in multimodal mathematics.  
 665 *arXiv:2501.04686*, 2025.  
 666

667 Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document  
 668 images. In *WACV*, 2021.  
 669

670 Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, et al.  
 671 MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement  
 672 learning. *arXiv:2503.07365*, 2025.  
 673

674 Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone.  
 675 Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of  
 676 Machine Learning Research*, 21(181):1–50, 2020.  
 677

678 OpenAI. Introducing OpenAI o1. <https://openai.com/o1/>, 2024.  
 679

680 OpenAI. Gpt-5 system card. [openai.com/index/gpt-5-system-card](https://openai.com/index/gpt-5-system-card), 2025.  
 681

682 YingZhe Peng, Gongrui Zhang, Xin Geng, and Xu Yang. LMM-R1. <https://github.com/TideDra/lmm-r1>, 2025.  
 683

684 Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language under-  
 685 standing by generative pre-training. *OpenAI*, 2018.  
 686

687 Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov,  
 688 Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al.  
 689 A generalist agent. *arXiv:2205.06175*, 2022.  
 690

691 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy  
 692 optimization algorithms. *arXiv:1707.06347*, 2017.  
 693

694 Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei  
 695 Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi,  
 696 Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in  
 697 rlrv. <https://rethink-rlrv.notion.site/Spurious-Rewards-Rethinking-g-Training-Signals-in-RLVR-1f4df34dac1880948858f95aeb88872f>, 2025.  
 698 Notion Blog.  
 699

700 Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and  
 701 Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large  
 702 language models. *arXiv:2406.17294*, 2024.  
 703

704 Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shang-  
 705 hang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning. *arXiv:2503.20752*,  
 706 2025.  
 707

708 Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2023.  
 709

702 Tencent Hunyuan3D Team. Hunyuan3D 2.0: Scaling diffusion models for high-resolution textured  
 703 3d assets generation. *arXiv:2501.12202*, 2025.

704

705 Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan,  
 706 et al. LLaMAV-01: Rethinking step-by-step visual reasoning in LLMs. *arXiv:2501.06186*, 2025.

707

708 Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha  
 709 Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully  
 710 open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing  
 711 Systems*, 37:87310–87356, 2024a.

712

713 Shengbang Tong, Zhuang Liu, Yuexiang Zhu, Xingjian Chen, Ruoyu Zhang, Bo Li, et al. Eyes wide  
 714 shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*,  
 2024b.

715

716 Bas Van Opheusden, Ionatan Kuperwajs, Gianni Galbiati, Zahy Bnaya, Yunqi Li, and Wei Ji Ma.  
 717 Expertise increases planning depth in human gameplay. *Nature*, 2023.

718

719 Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma,  
 720 Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust  
 721 multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

722

723 Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and  
 724 Hongsheng Li. Measuring multimodal mathematical reasoning with Math-Vision dataset. In  
 725 *NeurIPS*, 2024b.

726

727 Nicholas R. Waytowich, Devin White, M.D. Sunbeam, and Vinicius G. Goecks. Atari-GPT: Investi-  
 728 gating the capabilities of multimodal large language models as low-level policies for atari games.  
 729 *arXiv:2408.15950*, 2024.

730

731 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny  
 732 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*,  
 733 2022.

734

735 X.AI. Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>, 2024.

736

737 Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, et al. AtomThink:  
 738 A slow thinking framework for multimodal mathematical reasoning. *arXiv:2411.11930*, 2024.

739

740 Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu,  
 741 Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement  
 742 learning. *arXiv preprint arXiv:2502.14768*, 2025.

743

744 Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, et al.  
 745 R1-OneVision: Advancing generalized multimodal reasoning through cross-modal formalization.  
 746 *arXiv:2503.10615*, 2025.

747

748 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens,  
 749 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-  
 750 modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF  
 751 Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024a.

752

753 Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun,  
 754 Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal  
 755 understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

756

757 Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does  
 758 reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model?  
 759 *arXiv:2504.13837*, 2025.

760

761 Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, et al. MathVerse:  
 762 Does your multi-modal LLM truly see the diagrams in visual math problems? *arXiv:2403.14624*,  
 763 2024.

756 Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason  
757 without external rewards. *arXiv preprint arXiv:2505.19590*, 2025.  
758

759 Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh.  
760 R1-Zero’s “aha moment” in visual reasoning on a 2b non-sft model. *arXiv:2503.05132*, 2025.  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

810 APPENDIX  
811812 **Content**  
813

814 A. Data .....	16
815 A.1. Training Data Synthesis .....	16
816 A.2. Training Prompt in Visual Game Learning .....	17
817 A.3. Detail of Format Reward .....	19
818 B. Evaluation .....	20
819 B.1. Evaluation Detail of Atari Game .....	20
820 B.2. Ablation On Rotation Game .....	21
821 B.3. Synergistic Effects of Multi-Game Training .....	22
822 B.4. Auxiliary Task Selection: A Multi-Game Study .....	23
823 B.5. Reasoning Ability Boundary via Pass@ $k$ Evaluation .....	24
824 B.6. Detail of Evaluation Benchmarks .....	25
825 B.7. Evaluation on General Visual Capabilities .....	26
826 B.8. Inference Length Analysis .....	26
827 B.9. Reasoning Correlation Analysis Between Game and Math .....	27
828 C. Case Study .....	27
829	
830	
831	
832	
833	
834	
835	
836	
837	
838	
839	
840	
841	
842	
843	
844	
845	
846	
847	
848	
849	
850	
851	
852	
853	
854	
855	
856	
857	
858	
859	
860	
861	
862	
863	

864  
865

## A DATA

866  
867

## A.1 TRAINING DATA SYNTHESIS

868  
869  
870  
871  
872

Thanks to using the synthetic game data engine, we can flexibly generate large-scale training data with precisely controlled difficulty levels. This completely eliminates the need for extensive data filtering strategies used in previous rule-based RL work training on domain-specific data like math (Meng et al., 2025; Bae et al., 2025), where difficulty is hard to define and filtering can significantly reduce dataset size.

873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889

For the Snake game, the environment consists of a  $10 \times 10$  grid game board with two snakes of 1-grid initial length, where at each time step  $t$ , each snake receives one action to move, resulting in a new game state  $s_{t+1}$ . We define difficulty based on snake length—longer snakes create more complex game situations and more constrained movement options, closely aligning with how humans perceive difficulty when playing Snake. To generate meaningful moves that accomplish the objective of collecting more apples while remaining alive, we implement a policy network based on Proximal Policy Optimization (PPO) (Schulman et al., 2017). The observation space is represented as a  $10 \times 10$  grid with distinct values indicating empty cells (0), apples (1), the agent’s own body (2), and other agents’ bodies (3), stacked across 4 time steps to incorporate temporal information, resulting in an input tensor  $\mathbf{X} \in \mathbb{R}^{10 \times 10 \times 4}$ . The policy network architecture consists of two convolutional layers with  $3 \times 3$  kernels ( $C_1 = 16$  and  $C_2 = 32$  output channels), followed by fully connected layers that output action logits for the four possible movements, transformed into a probability distribution  $\pi(a|s)$  using softmax. To prevent suicidal moves, we incorporate action priors by masking logits for dangerous actions. The model employs the standard PPO objective with entropy regularization coefficient  $\beta = 0.01$ , value function coefficient  $\lambda = 0.5$ , and clipping parameter  $\varepsilon = 0.2$ . Agents receive rewards of  $r = +1$  for collecting apples and penalties of  $r = -1$  for dying, enabling them to learn complex behaviors such as obstacle avoidance, apple pursuit, and multi-step trajectory planning.

890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900

For the Rotation game, training data comprises synthetically generated visual puzzles focused on 3D spatial reasoning, utilizing 540 unique 3D object meshes (408 from Hunyuan3D 2.0 (Team, 2025) and 132 from Hunyuan3D 2.5). Our custom pipeline produces pairs of images ( $I_{\text{init}}, I_{\text{rot}}$ ) representing objects before and after defined rotations. Difficulty in Rotation is determined by the rotation angle between two images, where smaller angle differences present greater perceptual challenges. Each pair is generated through a precise sequence: establishing diverse initial viewpoints through compound transformations (base orientation plus additional z-axis rotation from  $\{0^\circ, 30^\circ, \dots, 330^\circ\}$  to prevent trivial pattern learning), then applying target rotations of either  $90^\circ$  or  $180^\circ$  exclusively around the z-axis. All objects are rendered at  $512 \times 512$  pixel resolution using a consistent perspective camera under standardized lighting conditions, resulting in approximately 32k unique pairs.

901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

Based on empirical results, we established optimal difficulty parameters for RL training across both games, which we ablate in Tab. 5c. This controlled progression of difficulty, made possible by our synthetic data generation approach, enables more effective learning trajectories compared to traditional data collection methods.

918 A.2 TRAINING PROMPT IN VISUAL GAME LEARNING  
919920 Prompt for Snake Game  
921922 Your role is to guide a snake within a Snake game featuring multiple apples.  
923924 This game is played on a board of size 10 by 10. The board uses a standard Cartesian coordinate system,  
925 where (0,0) represents the bottom-left position and (9,9) is the top-rightmost coordinate.  
926927 Apples at: {apple\_position}  
928929 Direction of Your Last Action: {last\_action}  
930931 Rules:  
932933 1) If you move onto an apple, you grow and gain 1 point.  
934 2) If your head moves to a position where its coordinates (x, y) are outside the board boundaries (meaning  
935 x < 0, x > 9, y < 0, or y > 9), or into a space occupied by another snake's body, or into a space occupied  
936 by your own body, you die. That's the worst move.  
937938 **3) The goal is to prioritize snake not die, then efficiently collecting apples. First avoid the worst  
939 move, then for each apple, find the nearest apple by calculating Manhattan distances. But only  
940 choose best next move to get closer the nearest apple if you can confirm best next move will not  
941 run outside the range of the listed coordinates, run into the position of another snake, or yourself.  
942 Otherwise it will be the worst move.**943 Your snake with the ID {snake\_id} in {snake\_color} has its head now positioned at {snake\_position}, and  
944 its body extends to {body\_position}. You should avoid your next move into your own snake's position.  
945946 Enemy snakes in {enemy\_color} positions: {enemy\_position}.  
947948 Decreasing your x coordinate is to the LEFT, increasing your x coordinate is to the RIGHT.  
949 Decreasing your y coordinate is DOWN, increasing your y coordinate is UP.  
950951 Read out another snake's position and apple position. Try to predict another snake's next move and avoid  
952 colliding with it.  
953954 Best answer is one of next move that is the closest to the apple and not lead to your death. Worst answer  
955 is all of next moves 1. makes your head's coordinates (x, y) are outside the board boundaries, meaning x  
956 < 0, x > 9, y < 0, or y > 9. 2. moves into a position occupied by another snake's body. 3. moves into a  
957 position occupied by body of yourself.  
958959 Check all the next moves to list out all the worst moves in <worst\_answer> tag. If no worst answer,  
960 return None for worst answer, e.g., "<worst\_answer>None</worst\_answer>"  
961962 The best answer and the worst answer are mutually exclusive and different.  
963964 You need first to give your reasoning process then to choose one of best next move and worst next move  
965 from ['UP', 'DOWN', 'LEFT', 'RIGHT'].  
966967 The reasoning process and answer are enclosed within <think> </think>, <best\_answer>  
968 </best\_answer> and <worst\_answer> </worst\_answer> tags, respectively, i.e., "<think> reasoning pro-  
969 cess here </think><best\_answer> one best move here </best\_answer><worst\_answer> all worst moves  
970 here </worst\_answer>"  
971

972  
973

## Prompt for Rotation Game

974  
975

I'm showing you 4 images. Images 1-2 are an example pair, and Images 3-4 are the test pair. In each pair, the first image shows the initial orientation, and the second shows the object after rotation.

976  
977

## ### EXAMPLE OF ROTATION ###

978  
979

Example: Image 1 shows the initial view and Image 2 shows the object after a 180 degree rotation.

980  
981  
982  
983

## ### YOUR TASK ###

Now, considering the transformation from Image 3 (initial) to Image 4 (rotated)

. Determine the angle of rotation from Image 3 to Image 4 on the plane

Analyze the rotation carefully using the example pair (Images 1-2) as a reference.

984  
985  
986  
987

## 1. Coordinate System Transformation:

- Draw an x-y coordinate system on both original and rotated images with origin at center
- Identify a distinct feature point and note its coordinates in both images
- Apply rotation matrix equations to verify the transformation

988  
989  
990

Example: A star icon at coordinates (3,1) in the original image appears at (-1,3) in the rotated image. Testing with the  $90^\circ$  clockwise rotation matrix  $[\cos(90^\circ), \sin(90^\circ); -\sin(90^\circ), \cos(90^\circ)]$  confirms the transformation from (3,1) to (-1,3), verifying a  $90^\circ$  clockwise rotation.

991

## 2. Angular Displacement Measurement:

- Mark the image center as the origin in both images
- Draw a straight line from center to a distinctive feature in both images
- Measure the angle between these two lines using counterclockwise as positive

995  
996  
997  
998

Example: A line from center to a red dot makes a  $30^\circ$  angle with horizontal in the original image. In the rotated image, this line makes a  $210^\circ$  angle with horizontal. The difference ( $180^\circ$ ) indicates a clockwise  $180^\circ$  rotation.

999  
1000  
1001  
1002

## 3. Symmetry Axis Tracking:

- Identify major symmetry axes in the original image
- Locate the same symmetry axes in the rotated image
- Calculate the angular displacement between original and rotated axes

1003  
1004  
1005

Example: A rectangular logo has vertical and horizontal symmetry axes. After rotation, the vertical axis now points right and horizontal points down. This  $90^\circ$  shift of both axes confirms a clockwise  $90^\circ$  rotation.

1006  
1007  
1008  
1009

## 4. Triangle Configuration Analysis:

- Select three non-collinear distinct points forming a triangle in both images
- Compare the orientation of this triangle in both images using vector cross products
- Determine rotation angle from the triangle's orientation change

1010  
1011  
1012  
1013

Example: Three points form a right triangle with vertices clockwise arranged. After rotation, the same triangle has its vertices arranged in counterclockwise order while maintaining the same shape. This inversion indicates a clockwise  $180^\circ$  rotation.

1014  
1015  
1016  
1017

## 5. Polar Coordinate Comparison:

- Convert key points to polar coordinates  $(r, \theta)$  relative to image center
- Compare  $\theta$  values of the same features in original and rotated images
- Calculate consistent angular difference across multiple points

1018  
1019  
1020

Example: A feature at polar angle  $45^\circ$  in the original image appears at  $135^\circ$  in the rotated image. Another feature shifts from  $10^\circ$  to  $100^\circ$ . Both show a  $+90^\circ$  shift in polar angle, confirming a clockwise  $90^\circ$  rotation.

1021  
1022

Choose the rotation angle from this list: ['counter clockwise 90', '180']

1023  
1024  
1025

The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags, respectively, i.e., "`<think>` reasoning process here `</think><answer>` answer here `</answer>`"

1026 While the model takes images as input to understand the current state of the game, we design  
 1027 a structural text prompt framework to also provide game guidance. Our game prompts  
 1028 consist of two parts: (1) game settings and (2) reasoning instructions. (1) To help the model under-  
 1029 stand the game environment, we describe the background, current game state, rules, goals,  
 1030 action space, *etc.* in text besides the input image. (2) In the reasoning instruction part,  
 1031 we provide specific thinking guidance since games can be approached with various thinking  
 1032 chains. To encourage broader thinking, we implement different types of reasoning instructions to  
 1033 guide decision-making process. Specifically, we used GPT-4o (Hurst et al., 2024) to synthesize  
 1034 mathematical thinking instructions for Snake, such as “finding the nearest apple by  
 1035 calculating Manhattan distances”, and spatial thinking instructions for Rotation, for  
 1036 example, “identify major symmetry axes in the original image”. With reasoning  
 1037 instructions for games, the obtained reasoning abilities generalize to downstream evaluation  
 1038 on visual math questions (Tab. 5a). **Bold text** indicates reasoning instructions synthesized by GPT-  
 1039 4o (Hurst et al., 2024).

1040 **A.3 DETAIL OF FORMAT REWARD**

1041 The format reward  $r_{\text{format}}$  validates whether the response follows the task-specific format:

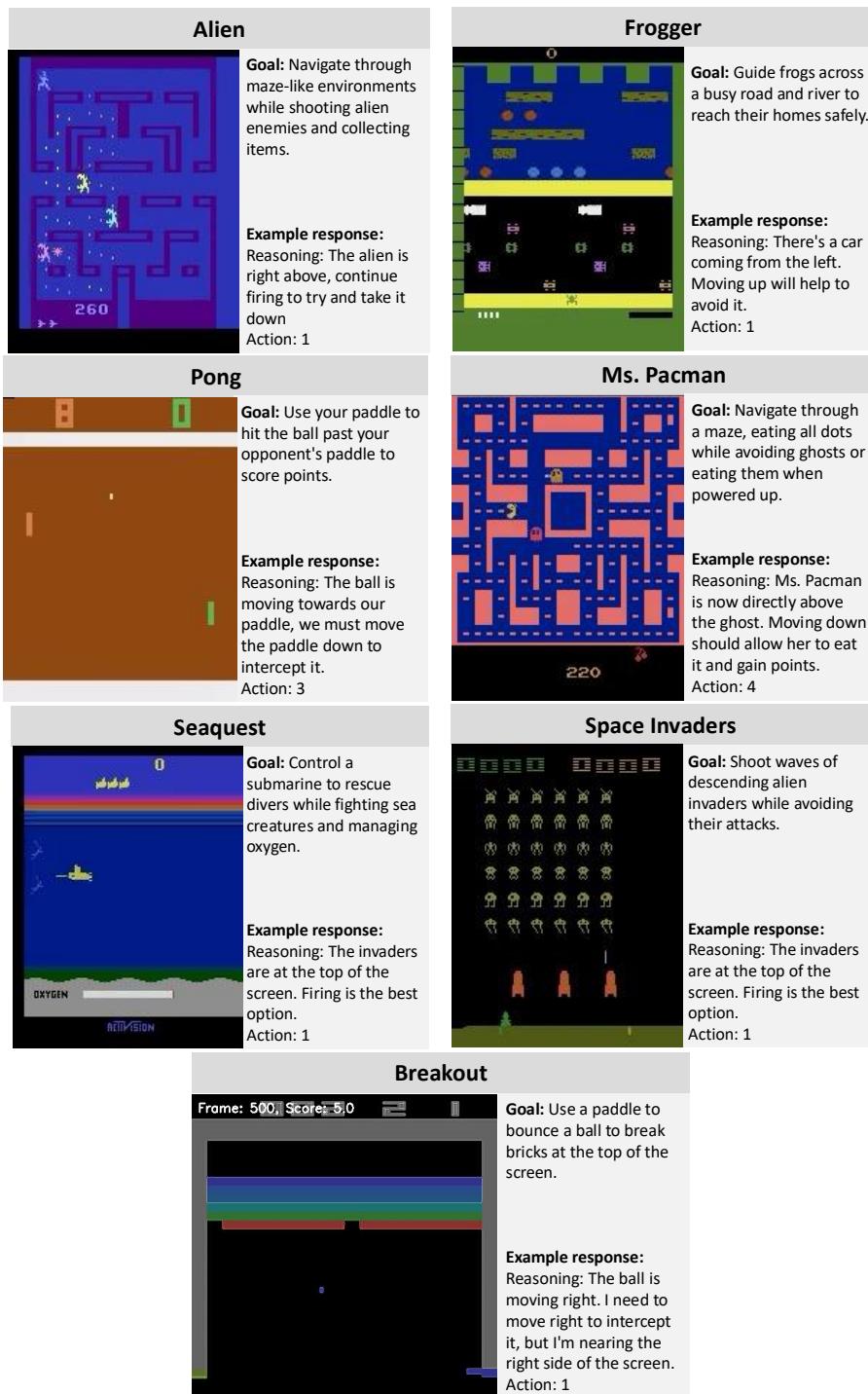
$$r_{\text{format}} = \begin{cases} 0.1, & \text{if the response follows the required format} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

1042 For Snake game, the desired format is:

1043 `<think>...</think><best_answer>...</best_answer><worst_answer>...</worst_answer>.`

1044 As suggested by the format, we encourage the model to predict both a positive move that moves  
 1045 toward the apple and a negative move that leads to failure. This reward encourages contrastive  
 1046 decision-making, which not only improves the model’s gameplay abilities but also boosts down-  
 1047 stream reasoning performance on visual math benchmarks. We ablate the effect in Tab. 5b. For the  
 1048 rotation task, the required format is simply `<think>...</think><answer>...</answer>`.

1049  
 1050  
 1051  
 1052  
 1053  
 1054  
 1055  
 1056  
 1057  
 1058  
 1059  
 1060  
 1061  
 1062  
 1063  
 1064  
 1065  
 1066  
 1067  
 1068  
 1069  
 1070  
 1071  
 1072  
 1073  
 1074  
 1075  
 1076  
 1077  
 1078  
 1079

1080 **B EVALUATION**  
10811082 **B.1 EVALUATION DETAIL OF ATARI GAME**  
10831130 **Figure 5: Goal and example response from model of Atari games used for evaluation.** We  
1131 implement 7 kinds of Atari games from Atari-GPT (Waytowich et al., 2024).1132  
1133

1134 To evaluate out-of-distribution generalization, we test ViGaL on Atari-GPT (Waytowich et al.,  
 1135 2024), a benchmark for evaluating MLLMs as decision-making agents in Atari video games, as  
 1136 shown in Fig. 5. The benchmark consists of seven different Atari games: Alien, Frogger, Pong, Ms.  
 1137 Pacman, Seaquest, Space Invaders and Breakout. These games present diverse visual environments  
 1138 which is different from Snake game and Rotation game, and require different strategic approaches to  
 1139 finish the goal, making them an ideal test bed for ViGaL evaluating out-of-distribution generalization  
 1140 capabilities.

1141 For evaluation, we input game frames as pixel observations to our model, following the established  
 1142 protocol in Atari-GPT. Specifically, each game frame is resized from  $210 \times 160 \times 3$  to  $512 \times 512 \times 3$ ,  
 1143 then provided to our model along with game-specific action information. We maintain a context  
 1144 buffer containing the two previous frames and responses together with the current frame to enable  
 1145 temporal reasoning. Following Atari-GPT, we implement frame skipping of 8 frames, which extends  
 1146 the standard 4-frame skipping in ALE to reduce computational intensity while preserving gameplay  
 1147 continuity.

1148 We evaluate our method through four independent rollouts of 1,000 timesteps each and report the  
 1149 average cumulative reward, with results presented in Tab. 1c.

1150

## 1151 B.2 ABLATION ON ROTATION GAME

1152

1153 **Table 6: Ablation study.** Similar to the evaluation in Tab. 5, we analyze how different aspects of  
 1154 our post-training strategy within the Rotation game affect downstream generalization benchmarks.  
 1155 The base model is Qwen2.5-VL-7B, with results shown in gray. The default settings from Tab. 2  
 1156 and Tab. 3 are highlighted in blue. We observe the same improvement trends for each strategy as  
 1157 reported in Tab. 5.

1158

prompt	(a) Prompt design.				(b) SFT vs. RL.				
	Avg.	Math	CLEVR+	Geo.	post-training	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8	base model	49.1	47.7	54.9	44.8
w/o Reasoning Instruction	61.4	48.9	80.4	54.8	SFT	55.6	44.0	75.4	47.5
w/ Reasoning Instruction	62.6	49.3	80.7	57.9	RL	62.6	49.3	80.7	57.9

	(c) Difficulty control.				
	difficulty control	Avg.	Math	CLEVR+	Geo.
	base model	49.1	47.7	54.9	44.8
	w/o difficulty control	61.0	48.0	80.2	54.8
	w/ difficulty control	62.6	49.3	80.7	57.9

1163

1164 As shown in Tab. 6, we conduct a similar ablation study to Tab. 5, but replace the Snake game  
 1165 environment with the Rotation game. Our results demonstrate the same consistent improvement  
 1166 trends on downstream generalization benchmarks for each strategy employed.

1167

1168 Specifically, we control the task difficulty by varying the rotation angles between two images. In  
 1169 the uncontrolled difficulty setting, the rotation angle between images can be clockwise  $90^\circ$ , counter-  
 1170 clockwise  $90^\circ$ , or  $180^\circ$ . However, we found that explicitly requiring the model to distinguish be-  
 1171 tween clockwise and counter-clockwise rotations leads to training difficulties. Therefore, we remove  
 1172 it and only retain option of clockwise  $90^\circ$  and  $180^\circ$  rotations.

1173

1174 Unlike the Snake game, we cannot conduct the ablations shown in Tab. 5e because the Rotation game  
 1175 is inherently vision-dependent and requires visual input. Similarly, we cannot perform the ablations  
 1176 in Tab. 5b because the Rotation game provides only binary answer options, making it impossible to  
 1177 meaningfully designate both "best" and "worst" answers simultaneously.

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

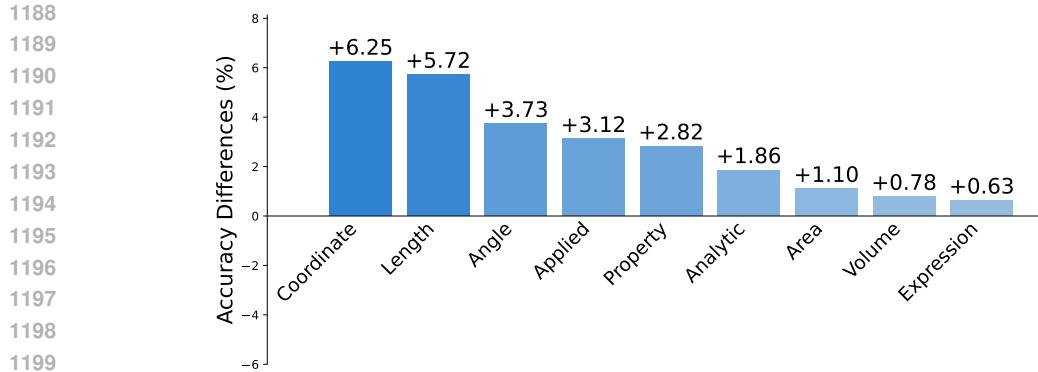


Figure 6: **Accuracy differences between ViGaL-Snake+Rotation and base model without RL training across mathematical subfields in Mathverse.** The synergistic effects of jointly training on two games observed suggest that complementary games can enhance overall mathematical reasoning capabilities.

### B.3 SYNERGISTIC EFFECTS OF MULTI-GAME TRAINING

As discussed in Sec. 3.1, our analysis reveals that each game develops distinct reasoning abilities in the model. To investigate potential combined benefits, we conducted experiments where models were trained simultaneously on *both* the Snake and Rotation games. Fig. 6 shows that joint training effectively combines the strengths of each individual game, improving performance across the mathematical areas where each game shows particular effectiveness, resulting in greater overall gains on Mathverse. These results suggest that strategically combining games with complementary strengths offers a simple yet effective approach to enhance model generalization abilities.

### B.4 GENERALIZATION BEYOND SNAKE AND ROTATION: A MULTI-GAME STUDY

A key question is whether the observed benefits are specific to Snake and Rotation, or whether they generalize across diverse game mechanics. To investigate this, we extend our study to four additional games: Sokoban, Maze, Tetris, and Sudoku. Sokoban requires pushing boxes to target positions while navigating spatial constraints; Maze involves pathfinding through obstacles to reach a goal; Tetris requires rotating and positioning falling shapes to complete rows; Sudoku demands filling grids while satisfying row, column, and box constraints. These games introduce cognitive challenges beyond Snake and Rotation, including strategic planning and logical deduction. This extended study allows us to examine whether the transfer of reasoning skills from gameplay to mathematical tasks is a general phenomenon or an artifact of our initial game selection.

**Consistent improvements across diverse games.** Tab. 7 shows that all games yield consistent gains across mathematical reasoning benchmarks, with improvements ranging from 1.5% to 1.7% on average. This demonstrates that developing transferable reasoning skills from visual gameplay is robust across diverse game mechanics and not specific to Snake and Rotation.

**Systematic transfer patterns through subject-level analysis.** To understand how different game mechanics transfer to specific mathematical sub-skills, we analyze performance on MathVerse using its official subject split, which assigns each problem to one of three subjects: Plane Geometry (angles and lengths), Functions (coordinate graphs and related diagrams), and Solid Geometry (volume and surface area). We use this subject-level split instead of the finer subfields because many of those subfields depend on overlapping skills. For example, the Coordinate, Property, and Expression subfields all require interpreting coordinate systems. The subject-level grouping therefore offers a more stable and disentangled proxy for distinct mathematical reasoning abilities, for instance by placing all questions that require interpreting coordinate systems under Functions.

Tab. 8 reports performance gains for all six games separately for each subject. Based on these performance profiles, we grouped the six games into three clusters using K-Means clustering to identify systematic transfer patterns:

1242 Table 7: Comparison of different auxiliary games on mathematical reasoning benchmarks. All  
 1243 models use Qwen2.5-VL-7B as the base. All games yield consistent improvements, demonstrating  
 1244 the generalizability of game-based transfer learning.

Model	Math Avg.	MathVerse	MathVista	MathVision
Qwen2.5-VL-7B	47.7	49.0	68.0	26.0
+ Maze	49.3	50.9	70.3	26.7
+ Rotation	49.3	50.4	71.2	26.3
+ Snake	49.4	51.1	70.7	26.5
+ Sokoban	49.3	50.5	70.1	27.2
+ Sudoku	49.3	51.2	69.0	27.6
+ Tetris	49.2	50.8	69.4	27.4

1258 Table 8: Performance gains on MathVerse subjects for each auxiliary game, showing accuracy im-  
 1259 provement relative to the Qwen2.5-VL-7B baseline. We use the official subject split from Math-  
 1260 Verse: Plane Geometry (angles and lengths), Functions (coordinate graphs and related diagrams),  
 1261 and Solid Geometry (volume and surface area). This subject-level grouping provides a more stable  
 1262 proxy for distinct mathematical reasoning abilities compared to finer subfields that depend on over-  
 1263 lapping skills.

Model	Plane Geometry	Functions	Solid Geometry
Baseline (Qwen2.5-VL-7B)	52.43	49.81	33.45
+Snake	+1.81	+3.77	+0.84
+Rotation	+2.32	+1.13	+0.50
+Maze	+1.96	+1.76	+1.84
+Tetris	+1.73	+2.01	+1.34
+Sokoban	+1.49	+1.89	+0.84
+Sudoku	+2.28	+2.27	+1.34

- **Cluster 1 (Snake, Sokoban):** Strongest transfer to **Functions** (+2.83% avg). Snake and Sokoban require tracking multiple objects moving in coordinate space: the snake’s growing body segments or boxes being pushed to target positions. Players must reason about how coordinate relationships between objects evolve as actions unfold, mirroring the core skill in Functions problems, understanding coordinate transformations (e.g., translations, scaling) and how values change under sequential operations.
- **Cluster 2 (Rotation):** Strongest transfer to **Plane Geometry** (+2.32%). This puzzle requires recognizing 3D object rotation angles, directly engaging angle reasoning and spatial relationships fundamental to plane geometry.
- **Cluster 3 (Maze, Sudoku, Tetris):** Balanced improvements across all three subjects. These games engage distinct reasoning modes: Maze requires navigating to find optimal paths; Sudoku demands symbolic constraint satisfaction without spatial movement; Tetris involves shape transformations and pattern completion. This diversity of non-overlapping cognitive demands supports reasoning across mathematical domains more broadly rather than specializing to a single subject.

1293 These findings reveal systematic relationships between game mechanics and mathematical reasoning  
 1294 categories. Beyond confirming the generalizability of our approach, the results provide practical  
 1295 guidance for auxiliary task selection: to improve a specific downstream skill, one can select existing  
 games whose cognitive requirements align with the target skill. Our framework offers a method for

1296 systematically choosing and combining off-the-shelf games based on observed cognitive transfer  
 1297 patterns.  
 1298

1299 **B.5 REASONING ABILITY BOUNDARY VIA PASS@ $k$  EVALUATION**  
 1300

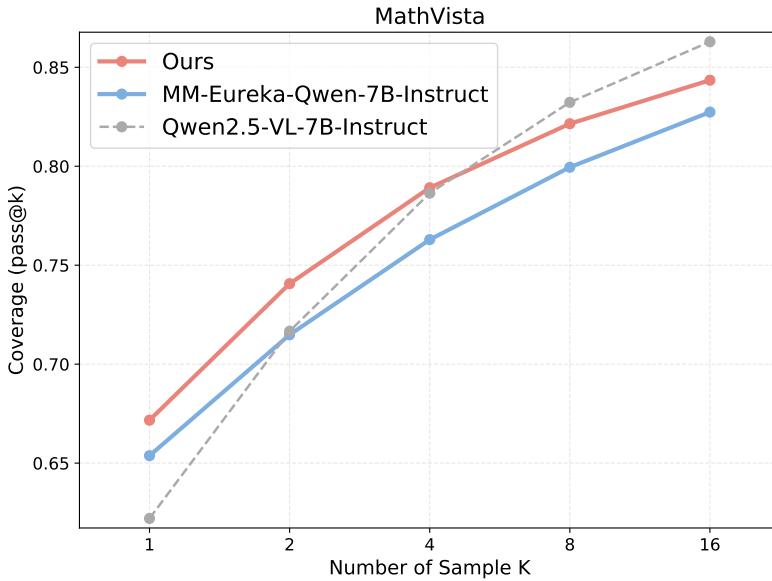


Figure 7: Pass@ $k$  performance curves on MathVista comparing base models with their zero-RL counterparts trained on mathematical data and game data, respectively.

We explore the reasoning ability boundary of models trained with different RL approaches by evaluating the pass@ $k$  metric. This metric measures the probability that at least one of  $k$  independent model samples solves a given problem, indicating the true scope or boundary of a model’s reasoning capability - essentially what problems the model can potentially solve given enough sampling attempts.

We evaluate the pass@ $k$  performance of three models: the Base Model without RL training, MM-Eureka-Qwen-7B-Instruct, and our ViGaL. As shown in Fig. 7, our ViGaL consistently demonstrates increasing pass@ $k$  scores on Mathverse as  $k$  increases. This finding suggests that our approach can effectively solve complex problems when allowed multiple reasoning attempts, uncovering capabilities not apparent in single-sample evaluations.

Moreover, compared to the other RL-trained model, MM-Eureka-Qwen-7B-Instruct, our model achieves a steeper improvement in pass@ $k$  as  $k$  increases. This indicates that ViGaL possesses a broader reasoning boundary and stronger reasoning abilities, enabling it to solve a wider range of problems when given sufficient opportunities to explore different solution paths.

Finally, our results demonstrate that as  $k$  increases, base models without RL training eventually outperform RL-trained models. This aligns with the findings in (Yue et al., 2025) that highlight a fundamental limitation of reinforcement learning with verifiable rewards (RLVR): while RL training significantly improves performance at small  $k$  values (e.g., pass@1), base models possess a wider coverage of solvable problems. This suggests a trade-off where RL optimization focuses on solving high-probability problems at the expense of broader solution coverage. Future work should explore RLVR algorithms that can improve pass@ $k$  performance across all values of  $k$ , effectively extending the reasoning boundary beyond that of the base model.

**B.6 DETAIL OF EVALUATION BENCHMARKS**

To obtain a clearer picture of the various facets of MLLM performance, we follow prior studies (Tong et al., 2024a; Li et al., 2024c) and systematically and carefully divide existing benchmarks

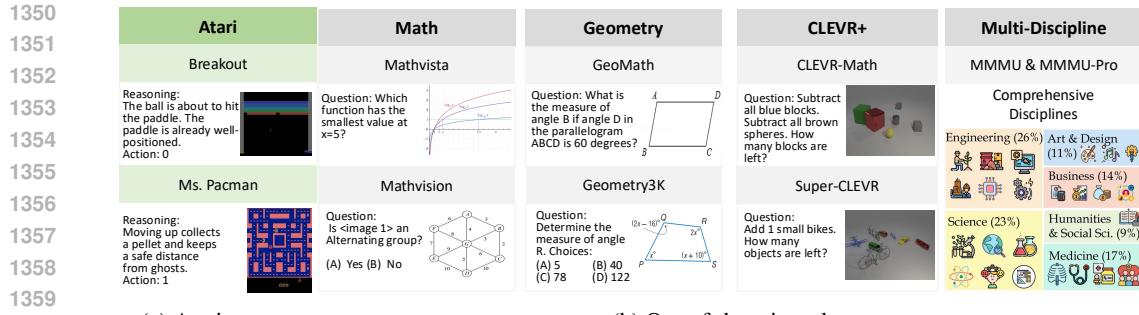


Figure 8: **Samples from our generalization reasoning benchmarks.** We evaluate the proposed Vi-GaL with two types of generalization: (a) *out-of-distribution* generalization, where models trained on our visual games are tested on unseen Atari games (Waytowich et al., 2024); and (b) *out-of-domain* generalization, where models trained only on game tasks are evaluated on diverse multimodal reasoning tasks including mathematical reasoning, geometric problem-solving, 3D understanding on CLEVR+ and multi-discipline reasoning on MMMU series.

into two broad groups: (i) *reasoning-oriented benchmarks*, which require multi-step or mathematical reasoning to solve the problems, and (ii) *general-purpose perception benchmarks*, which primarily assess broad visual understanding and perception abilities.

For reasoning-oriented benchmarks, we comprehensively evaluate the visual reasoning generalization capabilities of RL through gaming on a diverse collection of tasks that specifically demand advanced visual reasoning skills, including math-focused tasks like Math and Geometry, and other comprehensive reasoning benchmarks beyond math, like CLEVR+ and Multi-Discipline. Fig. 8b illustrates specific examples from each benchmark.

- **Math** evaluates multimodal math reasoning with widely-used datasets: MathVista (test-mini) (Lu et al., 2024), MathVerse (testmini) (Zhang et al., 2024), and MathVision (test) (Wang et al., 2024b). MathVista offers diverse problems spanning VQA, logic, algebra, and geometry; MathVerse emphasizes algebraic and geometric image comprehension; MathVision tests abstract visual reasoning.
- **Geometry** evaluates structural interpretation skills across mathematical diagrams, medical images, charts, and architectural layouts. It uses datasets GeoMath (Geo170K (Gao et al., 2023), Math360K (Shi et al., 2024)) and Geometry3K (Lu et al., 2021), featuring both choice and non-choice questions. Following Reason-RFT (Tan et al., 2025), we test with 820 GeoMath and 800 Geometry3K samples.
- **CLEVR+** evaluates the integration of mathematical and spatial reasoning skills through challenging arithmetic problems in complex 3D block-based scenes, including sub-tasks on CLEVR-Math (Lindström & Abraham, 2022) and Super-CLEVR (Li et al., 2023). Following Reason-RFT (Tan et al., 2025), we use 1K test samples from each of CLEVR-Math and Super-CLEVR.
- **Multi-Discipline** evaluates college-level expert knowledge across six disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. We follow the evaluation setting of MMMU (Yue et al., 2024a) val set (900 questions) and MMMU-Pro (Yue et al., 2024b) overall score (average of standard 10-option and vision-only settings).

For general-purpose perception benchmarks, we systematically evaluate comprehensive visual capabilities. Following previous work, these benchmarks are categorized into three distinct types: General, Vision-Centric, and OCR & Chart.

- **General** benchmarks assess fundamental visual understanding capabilities. We evaluate MuirBench (Wang et al., 2024a) for multi-image understanding and CRPE (Kazemzadeh et al., 2014) for relation understanding.

1404	Model	General				Vision-Centric					OCR & Chart					
		Avg.	Avg.	Muir-Bench	CRPE <sub>rel</sub>	Avg.	MMVP	Real-WorldQA	MMStar	BLINK <sub>val</sub>	MME <sub>p</sub>	Avg.	AI2D <sub>w.M.</sub>	SEED-Bench-2+	DocVQA	OCR-Bench
Proprietary Model																
1407	GPT-4o (Hurst et al., 2024)	74.8	72.3	68.0	76.6	69.4	—	75.4	64.7	68.0	1614	82.6	84.6	72.0	91.1	736
General Multimodal Language Model																
1409	Qwen2.5-VL-7B (Bai et al., 2023)	72.4	68.0	59.6	76.4	65.8	74.3	68.5	63.9	56.4	1698	83.3	83.9	70.4	95.7	864
Multimodal Reasoning Model Post-Trained on Qwen2.5-VL-7B																
1410	R1-Onenision-7B (Yang et al., 2025)	—	66.8	46.3	87.3	56.5	61.3	58.0	57.8	48.7	1504	—	—	—	—	—
1411	R1-VL-7B (Chen et al., 2025b)	67.4	63.3	54.1	72.4	59.6	70.3	61.4	55.6	51.0	1657	79.2	81.7	66.4	89.4	81.0
1412	MM-Eureka-Qwen-7B (Meng et al., 2025)	71.8	<b>68.9</b>	61.1	76.7	65.1	74.3	66.1	65.9	54.0	1626	81.5	84.3	68.2	92.0	87.0
1413	Reason-RFT-Zero-7B (Tan et al., 2025)	68.4	66.9	58.5	75.2	58.5	58.0	65.3	59.1	51.6	1653	79.8	83.3	68.0	88.1	82.0
1414	VLAA-Thinker-7B (Chen et al., 2025a)	69.7	65.9	57.1	74.6	62.6	71.6	65.4	60.4	53.0	1593	80.6	83.4	67.4	90.9	84.5
1415	OpenVLThinker-7B (Deng et al., 2025)	—	64.3	52.8	75.8	50.4	32.3	60.2	59.1	49.9	1513	—	—	—	—	—
1416	ViGaL Snake + Rotation	72.2	68.6	60.5	76.7	<b>65.7</b>	74.6	67.3	65.4	55.6	1685	<b>82.2</b>	84.8	69.1	92.7	86.6

Table 9: **Main results on multimodal language benchmarks targeting more general and comprehensive visual ability.** We compare with models post-trained on Qwen2.5-VL-7B (Bai et al., 2023). Best category averages are highlighted in **bold**. Note that MME<sub>p</sub> is excluded from vision-centric category average accuracy due to scale differences.

- **Vision-Centric** benchmarks thoroughly evaluate perception, real-world understanding, and multi-modal capabilities. We assess MMVP (Tong et al., 2024b), RealWorldQA (X.AI, 2024), MMStar (Chen et al., 2024a), MME (Fu et al., 2023), and BLINK (Fu et al., 2024).
- **OCR & Chart** understanding benchmarks focus on text-rich visual content. We specifically use AI2D (Kembhavi et al., 2016) for diagram understanding, SEED-Bench-2-Plus (Li et al., 2024a) for text-rich visual comprehension, DocVQA (Mathew et al., 2021) for document understanding, and OCRBench (Liu et al., 2023) for comprehensive OCR evaluation.

## B.7 EVALUATION ON GENERAL VISUAL CAPABILITIES

As discussed in Sec. 3.1, we have already demonstrated that our game RL training can generalize to visual math reasoning and multi-discipline reasoning benchmarks. To evaluate whether reasoning improvements come at the cost of general visual understanding, we test ViGaL Snake + Rotation on diverse general visual benchmarks. The results, summarized in Tab. 9, show that our model preserves general visual performance at a level comparable to the Qwen2.5-VL-7B base model, while delivering stronger gains in mathematical reasoning. In contrast, prior RL-based approaches often sacrifice visual accuracy when optimized for reasoning. This confirms that our gameplay-driven strategy enhances reasoning ability without compromising broader visual competencies.

## B.8 INFERENCE LENGTH ANALYSIS

Recent reinforcement learning studies (Xie et al., 2025; Aggarwal & Welleck, 2025) have raised questions about whether performance improvements stem from genuinely enhanced reasoning capabilities or merely from models generating longer responses. To address this concern, we analyze the relationship between response length and performance for models trained with our game-based approach.

Model	Response Length	Math Avg.
Qwen2.5-VL-7B (baseline)	250	47.7
ViGaL (ours, RL on games)	268	<b>50.6</b>

Table 10: Response length and performance on visual math benchmarks. Our game-based RL approach achieves significant performance gains while maintaining comparable inference costs.

Table 10 demonstrates that our performance improvements are not simply due to increased verbosity. Our ViGaL model achieves substantial performance gains (50.6% vs. 47.7%) while maintaining nearly identical inference costs—the response length increases by only 7% (268 vs. 250 tokens). This minimal increase in response length, coupled with the significant accuracy improvement, indicates that the model has learned transferable skills rather than merely generating longer outputs.

1458 These results suggest that game-based RL training enables effective knowledge transfer from game  
 1459 environments to mathematical problem-solving. For example, spatial reasoning skills acquired from  
 1460 the Rotation game and coordinate recognition abilities developed through Snake gameplay transfer  
 1461 effectively to visual math tasks. The model thus learns genuine problem-solving strategies while  
 1462 maintaining inference costs.

1463

## 1464 B.9 REASONING CORRELATION ANALYSIS BETWEEN GAME AND MATH

1465

1466 To understand the mathematical reasoning patterns in snake game playing, we developed a system-  
 1467 atic approach to extract and analyze reasoning steps from multiple gameplay traces. Our methodol-  
 1468 ogy uses GPT-5 OpenAI (2025) as an analytical tool in a two-stage process.

1469

1470 In Stage A, we collect multiple snake game "thinking traces", which are detailed reasoning se-  
 1471 quences generated during gameplay, and distill them into a generalized set of 8 core reasoning steps.  
 1472 These steps abstract away specific details like exact coordinates or particular board configurations to  
 1473 capture fundamental cognitive operations. The operations include parsing board state, enumerating  
 1474 moves, safety screening, path metric selection, distance computation, target identification, enemy  
 1475 anticipation, and move ranking. This summarization ensures our analysis focuses on transferable  
 1476 reasoning patterns rather than game-specific instances.

1477

1478 In Stage B, we quantify how mathematical each reasoning step is by evaluating its correlation with  
 1479 nine distinct mathematical aspects. We use a simple 3-level scoring system where 0 means no cor-  
 1480 relation, 1 means low correlation, and 2 means high correlation. GPT-5 analyzes how strongly each  
 1481 step relates to mathematical concepts such as coordinate manipulation, distance metrics, analytical  
 1482 reasoning, and geometric properties.

1483

1484 The resulting correlation matrix in Tab. 11 reveals clear patterns. Coordinate-based reasoning domi-  
 1485 nates steps that involve spatial parsing and movement planning, particularly Steps 1 through 3, Step  
 1486 5, and Step 7. Meanwhile, analytical and length-based reasoning become prominent in optimiza-  
 1487 tion steps like target identification and move ranking, seen in Steps 6 and 8. Steps 4 and 5, which involve  
 1488 path metrics and distance computation, show high correlation with both coordinate systems and  
 1489 length calculations. This confirms the geometric nature of pathfinding in grid-based environments.  
 1490 Our systematic analysis demonstrates that even seemingly simple game-playing behaviors require  
 1491 sophisticated integration of multiple mathematical reasoning capabilities.

1492

### Prompt Template for Reasoning Step Extraction and Correlation Analysis

1493

#### Stage A - Step Extraction:

1494 Given multiple snake game thinking traces, extract N general reasoning steps (6-9 steps) that capture the  
 1495 core operations. Abstract away instance-specific details and output:

1496

- 1497 • Short, action-oriented step names with one-line descriptions
- 1498 • General patterns covering: state parsing, move generation, safety screening, target selection via  
 1499 distance, opponent awareness, scoring/tie-breaks, decision, reporting

1500

#### Stage B - Mathematical Aspect Correlation:

1501

1502 For each extracted step, assign correlation levels (0/1/2) to these mathematical aspects:

1503

- 1504 • **Expression:** Formatting/structuring outputs
- 1505 • **Coordinate:** Reading/writing positions, mapping moves to (x,y)
- 1506 • **Area:** Board regions/bounds as areas
- 1507 • **Volume:** 3D spatial reasoning (if applicable)
- 1508 • **Applied:** Goal-directed task execution
- 1509 • **Property:** Rules/invariants (bounds, occupancy, collision)
- 1510 • **Angle:** Angle-based path reasoning
- 1511 • **Analytic:** Selection/optimization, tie-break logic
- 1512 • **Length:** Distance metrics (Manhattan/L1, grid paths)

1513 Output as structured table with integer scores only (0 = no correlation, 1 = low, 2 = high).

1512  
 1513 Table 11: Correlation Matrix of each step reasoning trace of playing snake game with solving math  
 1514 questions. (0=No Correlation, 1=Low Correlation, 2=High Correlation)

Step	Operation	Expression	Coordinate	Area	Volume	Applied	Property	Angle	Analytic	Length
1	Parse the board state									
2	Enumerate candidate moves									
3	Safety screening (worst-move test)									
4	Choose a path metric									
5	Compute distances to apples									
6	Identify nearest target									
7	Anticipate enemy motion (when present)									
8	Rank safe moves									

1515  
 1516 **Correlation Legend:**  
 1517

1518 = No Correlation      = Low Correlation      = High Correlation

1566 C CASE STUDY  
1567

**Rotation**

Question:  
As shown in the figure, the straight line  $a$  and the straight line  $b$  are intercepted by the straight line  $c$ . Line  $b$  is perpendicular to  $c$ , and the foot of the perpendicular is point  $A$ .  $\angle 1 = 70^\circ$ . If line  $b$  is parallel to line  $a$ , how many degrees clockwise can line  $b$  be rotated around point  $A$ ?  
Choices: A:  $70^\circ$  B:  $50^\circ$  C:  $30^\circ$  D:  $20^\circ$

**Qwen2.5-VL-7B-Instruct**

- ✓ 1. Identify the given information: Line  $b$  is perpendicular to line  $c$ . The foot of the perpendicular is point  $A$ . Angle  $1$  is  $70^\circ$ . Line  $b$  is parallel to line  $a$ .
- ✗ 2. Understand the geometric relationships: Since line  $b$  is perpendicular to line  $c$ , angle  $1$  is the angle between line  $a$  and line  $c$ . Since line  $b$  is parallel to line  $a$ , the angle between line  $b$  and line  $c$  is the same as the angle between line  $a$  and line  $c$ , which is  $70^\circ$ .
- ✗ 3. Determine the angle of rotation: The angle of rotation is the difference between the initial angle  $70^\circ$ , and the final angle  $0^\circ$ , since line  $b$  is perpendicular to line  $c$ .
- ✗ Therefore, the angle of rotation is  $70^\circ$ . Thus, the line  $b$  can be rotated  $70^\circ$  clockwise around the point  $A$ . The correct answer is: A

**Ours**

- ✓ 1. Identify the given information: Line  $b$  is perpendicular to line  $c$ . The foot of the perpendicular is point  $A$ .  $\angle 1 = 70^\circ$ . Line  $b$  is parallel to line  $a$ .
- ✓ 2. Analyze the geometric relationships: Since line  $b$  is perpendicular to line  $c$ ,  $\angle BAC = 90^\circ$ . Line  $b$  is parallel to line  $a$ , so the corresponding angles are equal. The angle  $\angle 1$  is given as  $70^\circ$ .
- ✓ 3. Determine the rotation angle: - The angle  $\angle 1$  represents the angle between line  $b$  and the vertical line  $c$ . Since line  $b$  is perpendicular to line  $c$ , the angle between line  $b$  and line  $a$  (which is parallel to line  $b$ ) is  $90^\circ - 70^\circ = 20^\circ$ .
- ✓ Therefore, the line  $b$  can be rotated  $20^\circ$  clockwise around the point  $A$ . The correct answer is: D.

(a) A case study from Mathverse. Base model misinterpreted the geometric configuration and rotation direction, while our model correctly identified the perpendicular relationship and calculated the proper angle.

**Coordinate**

Question:  
Determine whether the graphed function is even, odd, or neither.

**Qwen2.5-VL-7B-Instruct**

- ✗ Let's analyze the graph:  
1. Symmetry with respect to the y-axis: If the graph were symmetric with respect to the y-axis, then for every point  $(x, y)$  on the graph, the point  $(-x, y)$  would also be on the graph. However, the graph in the image does not appear to be symmetric with respect to the y-axis. For example, the point  $(3, y)$  does not have a corresponding point  $(-3, y)$  on the graph.
- ✓ 2. Symmetry with respect to the origin: If the graph were symmetric with respect to the origin, then for every point  $(x, y)$  on the graph, the point  $(-x, -y)$  would also be on the graph. However, the graph in the image does not appear to be symmetric with respect to the origin. For example, the point  $(3, y)$  does not have a corresponding point  $(-3, -y)$  on the graph.
- ✗ Since the graph does not exhibit symmetry with respect to either the y-axis or the origin, the function is neither even nor odd. The final answer is: neither.

**Ours**

- ✓ Let's examine the graph of the function. The graph shows a parabola that opens upwards and is symmetric about the y-axis. This symmetry indicates that the function is even.
- ✓ To confirm, we can check a few points:  
For  $x = 2$ , the function value is  $f(2)$ . For  $x = -2$ , the function value is  $f(-2)$ . Since the graph is symmetric about the y-axis,  $f(-2) = f(2)$ . This confirms that  $f(-x) = f(x)$ , which is the definition of an even function.
- ✓ Therefore, the function is even. The final answer is: even.

(b) A case study from Mathverse. Base model misperceived critical visual information like symmetry and coordinates in graphs, while our model demonstrated accurate visual perception for mathematical elements.

1601 Figure 9: Comparison of base model and our model after rule-based RL training, showing improved  
1602 visual-mathematical reasoning on geometric and coordinate problems.  
1603

1604 We provide quantitative comparison examples below to demonstrate reasoning improvements on  
1605 mathematical problems after RL training. In Fig. 9a, when solving a geometric angle problem, the  
1606 base model fails to correctly interpret the critical relationship between perpendicular lines and corre-  
1607 sponding angles. It makes contradictory assumptions about angle measures, leading to an incorrect  
1608 calculation of the required rotation. In contrast, our ViGaL precisely tracks the geometric constraints  
1609 and properly calculates the angle difference between initial and target positions. In Fig. 9b, when  
1610 analyzing function properties from a graph, the base model incorrectly claims the function lacks  
1611 symmetry despite clear visual evidence. It fails to recognize the fundamental y-axis symmetry of  
1612 the parabola shown in the image. Our model immediately identifies this critical symmetrical pattern  
1613 and correctly applies the appropriate mathematical definition of an even function, demonstrating  
1614 enhanced visual perception of mathematical structures.  
1615

1616 D USE OF LARGE LANGUAGE MODELS  
1617

1618 In accordance with ICLR 2026 policies, we disclose that a large language model (LLM) was used  
1619 during paper preparation. Specifically, LLMs were employed for grammar correction, wording

1620 refinement, and drafting of non-technical text passages. All research ideas, methods, experiments,  
1621 and analyses were conceived, implemented, and validated by the authors, who take full responsibility  
1622 for the correctness and integrity of the content.

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673