|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Step** | **Module** | **Block** | **IN shape** | **OUT shape** | **memory\_allocated** | | **max\_memory\_allocated** | **memory\_reserved** | **cuda context** | **nvidia-smi** | |  |  |  |  |  | **理论分析** | **实际消耗** | **实际消耗** | **实际消耗** | **实际消耗** | **实际消耗** | | 0 | Pre\_inputs | / | / | 2\*64\*64\*4  2\*77\*768  1 | **0.3 MiB, 295.1 KiB**  （2\*64\*64\*4+2\*77\*768+1）\*2/1024=295.1KiB | **0.3 MiB, 295.5 KiB** | **0.7 MiB, 757.5 KiB** | **2.0 MiB** | **620 MiB** | **622.0 MiB** | | 1 | Load\_weights | / | / | 845.6 MB | **1652.9 MiB, 1692562.0 KiB** | **1652.9 MiB, 1692562.0 KiB** | **1652.9 MiB, 1692562.0 KiB** | **1704.0 MiB** | **2324.0 MiB** | | 2 | Pre\_Timestep | Time\_emb | 2\*1280  2\*320 | 2\*1280  2\*320 | **1652.9 MiB，1692568.2 KiB**  (2\*1280+2\*320)\*2/1024=6.2KiB  1692562.0+6.2=1692568.2 KiB | **1652.9 MiB, 1692568.5 KiB** | **1657.9 MiB, 1697737.5 KiB** | **1724.0 MiB** | **2344.0 MiB** | | 3 | Pre\_process | Conv2D | 2\*320\*64\*64 | 2\*320\*64\*64 | **1657.9 MiB，1697688.2 KiB**  （2\*320\*64\*64)\*2/1024=5120 KiB  1692568.2+5120=1697688.2 KiB | **1657.9 MiB, 1697688.5 KiB** | **1667.9 MiB, 1707938.0 KiB** | **1724.0 MiB** | **2344.0 MiB** | | 4 | CrossAttnDownBlock2D-1 | ResnetBlock2D | 2\*320\*64\*64 | 2\*320\*64\*64 | **1662.9MiB，1702808.2 KiB**  （2\*320\*64\*64)\*2/1024=5120 KiB  1697688.2+5120=1702808.2 KiB | **1662.9 MiB, 1702808.5 KiB** | **1679.9 MiB, 1720218.0 KiB** | **1736.0 MiB** | **2356.0 MiB** | | 5 | Transformer | 2\*320\*64\*64 | 2\*320\*64\*64 | **1667.9 MiB，1707928.2 KiB**  （2\*320\*64\*64)\*2/1024=5120 KiB  1702808.2+5120=1707928.2 KiB | **1667.9 MiB, 1707928.5 KiB** | **3248.9 MiB, 3306392.5 KiB** | **3292.0 MiB** | **3912.0 MiB** | | 6 | ResnetBlock2D | 2\*320\*64\*64 | 2\*320\*64\*64  （覆盖，不申请内存） | **1667.9 MiB，1707928.2 KiB** | **1667.9 MiB, 1707928.5 KiB** | | 7 | Transformer | 2\*320\*64\*64 | 2\*320\*64\*64 | **1672.9 MiB，1713048.2 KiB**  （2\*320\*64\*64)\*2/1024=5120 KiB  1707928.2+5120=1713048.2 KiB | **1672.9 MiB, 1713048.5 KiB** | | 8 | DownBlock | 2\*320\*32\*32 | 2\*320\*32\*32 | **1674.1 MiB, 1714328.2 KiB**  （2\*320\*32\*32)\*2/1024=1280 KiB  1713048.2 +1280=1714328.2 KiB | **1674.1 MiB, 1714328.5 KiB** | |  | RES | 2\*320\*64\*64  2\*320\*64\*64  2\*320\*32\*32 | 2\*320\*64\*64  2\*320\*64\*64  2\*320\*32\*32 | **1669.1 MiB, 1709208.2 KiB**  (2\*320\*64\*64\*2+2\*320\*32\*32)\*2/1024=11520.0 KiB  1697688.2 +11520.0=1709208.2KiB | **1669.1 MiB, 1709208.5 KiB** | | 9 | CrossAttnDownBlock2D-2 | ResnetBlock2D | 2\*640\*32\*32 | 2\*640\*32\*32 | **1671.6 MiB, 1711768.2 KiB**  (2\*640\*32\*32)\*2/1024=2560 KiB  1709208.2+2560=1711768.2 KiB | **1671.6 MiB, 1711768.5 KiB** | | 10 | Transformer | 2\*640\*32\*32 | 2\*640\*32\*32 | **1674.1 MiB, 1714328.2 KiB**  (2\*640\*32\*32)\*2/1024=2560 KiB  1711768.2 +2560=1714328.2 KiB | **1674.1 MiB, 1714328.5 KiB** | | 11 | ResnetBlock2D | 2\*640\*32\*32 | 2\*640\*32\*32  （覆盖，不申请内存） | **1674.1 MiB, 1714328.2 KiB** | **1674.1 MiB, 1714328.5 KiB** | | 12 | Transformer | 2\*640\*32\*32 | 2\*640\*32\*32 | **1676.6 MiB, 1716888.2 KiB**  (2\*640\*32\*32)\*2/1024=2560 KiB  1714328.2+2560=1716888.2 KiB | **1676.6 MiB, 1716888.5 KiB** | | 13 | DownBlock | 2\*640\*16\*16 | 2\*640\*16\*16 | **1677.3 MiB, 1717528.2 KiB**  (2\*640\*16\*16)\*2/1024=640 KiB  1716888.2+640=1717528.2 KiB | **1677.3 MiB, 1717528.5 KiB** | |  | RES | 2\*640\*32\*32  2\*640\*32\*32  2\*640\*16\*16 | 2\*640\*32\*32  2\*640\*32\*32  2\*640\*16\*16 | **1674.8 MiB, 1714968.2 KiB**  (2\*640\*32\*32\*2+2\*640\*16\*16)\*2/1024=5760 KiB  1709208.2+5760=1714968.2 KiB | **1674.8 MiB, 1714968.5 KiB** | | 14 | CrossAttnDownBlock2D-3 | ResnetBlock2D | 2\*1280\*16\*16 | 2\*1280\*16\*16 | **1676.0 MiB, 1716248.2 KiB**  (2\*1280\*16\*16)\*2/1024=1280 KiB  1714968.2 +1280=1716248.2 KiB | **1676.0 MiB, 1716248.5 KiB** | | 15 | Transformer | 2\*1280\*16\*16 | 2\*1280\*16\*16 | **1677.3 MiB, 1717528.2 KiB**  (2\*1280\*16\*16)\*2/1024=1280 KiB  1716248.2+1280=1717528.2 KiB | **1677.3 MiB, 1717528.5 KiB** | | 16 | ResnetBlock2D | 2\*1280\*16\*16 | 2\*1280\*16\*16  （覆盖，不申请内存） | **1677.3 MiB, 1717528.2 KiB** | **1677.3 MiB, 1717528.5 KiB** | | 17 | Transformer | 2\*1280\*16\*16 | 2\*1280\*16\*16 | **1678.5 MiB, 1718808.2 KiB**  (2\*1280\*16\*16)\*2/1024=1280 KiB  1717528.2+1280=1718808.2 KiB | **1678.5 MiB, 1718808.5 KiB** | | 18 | DownBlock | 2\*1280\*8\*8 | 2\*1280\*8\*8 | **1678.8 MiB, 1719128.2 KiB**  (2\*1280\*8\*8)\*2=320 KiB  1718808.2+320=1719128.2 KiB | **1678.8 MiB, 1719128.5 KiB** | |  | RES | 2\*1280\*16\*16  2\*1280\*16\*16  2\*1280\*8\*8 | 2\*1280\*16\*16  2\*1280\*16\*16  2\*1280\*8\*8 | **1677.6 MiB, 1717848.2 KiB**  (2\*1280\*16\*16\*2+2\*1280\*8\*8)\*2/1024=2880 KiB  1714968.2+2880=1717848.2 KiB | **1677.6 MiB, 1717848.5 KiB** | | 19 | DownBlock2D | ResnetBlock2D | 2\*1280\*8\*8 | 2\*1280\*8\*8 | **1677.9 MiB, 1718168.2 KiB**  (2\*1280\*8\*8)\*2=320 KiB  1717848.2+320=17181628. 2 KiB | **1677.9 MiB, 1718168.5 KiB** | | ResnetBlock2D | 2\*1280\*8\*8 | 2\*1280\*8\*8 | **1678.2 MiB, 1718488.2 KiB**  (2\*1280\*8\*8)\*2=320 KiB  1718488.2+320= 1718488.2 KiB | **1678.2 MiB, 1718488.5 KiB** | | 20 | Mid2DCrossAttn | ResnetBlock2D | 2\*1280\*8\*8 | 2\*1280\*8\*8 | **1678.5 MiB, 1718808.2 KiB**  (2\*1280\*8\*8)\*2=320 KiB  1718488.2+320=1718808.2 KiB | **1678.5 MiB, 1718808.5 KiB** | | 21 | Transformer | 2\*1280\*8\*8 | 2\*1280\*8\*8  （覆盖，不申请内存） | **1678.5 MiB, 1718808.2 KiB** | **1678.5 MiB, 1718808.5 KiB** | | 22 | ResnetBlock2D | 2\*1280\*8\*8 | 2\*1280\*8\*8 | **1678.8 MiB, 1719128.2 KiB**  (2\*1280\*8\*8)\*2=320 KiB  1718808.2+320= 1719128.2 KiB | **1678.8 MiB, 1719128.5 KiB** | | 23 | UPblock2D | ResnetBlock2D | 2\*2560\*8\*8 | 2\*1280\*16\*16 | **1678.8 MiB, 1719128.2 KiB**  (2\*1280\*16\*16)\*2/1024=1280 KiB  1719128.2+1280=1719128.2 KiB | **1678.8 MiB, 1719128.5 KiB** | | 24 | ResnetBlock2D | 2\*2560\*8\*8 | 2\*1280\*16\*16  （覆盖，不申请内存） | **1719128.2+1280=1719128.2 KiB** | **1678.8 MiB, 1719128.5 KiB** | | 25 | ResnetBlock2D | 2\*2560\*8\*8 | 2\*1280\*16\*16  （覆盖，不申请内存） | **1719128.2+1280=1719128.2 KiB** | **1678.8 MiB, 1719128.5 KiB** | | 26 | Interpolate2D | 2\*1280\*16\*16 | 2\*1280\*16\*16 | **1680.1 MiB, 1720408.2 KiB**  (2\*1280\*16\*16)\*2/1024=1280 KiB  1719128.2+1024=1720408.2 KiB | **1680.1 MiB, 1720408.5 KiB** | | 27 | CrossAttnUPBlock2D-1 | ResnetBlock2D | 2\*1280\*16\*16  2\*1280\*16\*16 | 2\*1280\*16\*16 | **1681.0 MiB, 1721368.2 KiB** | **1681.0 MiB, 1721368.5 KiB** | | 28 | Transformer | 2\*1280\*16\*16 | 2\*1280\*16\*16  （覆盖，不申请内存） | **1681.0 MiB, 1721368.5 KiB** | **1681.0 MiB, 1721368.5 KiB** | | 29 | ResnetBlock2D | 2\*1280\*16\*16  2\*1280\*16\*16 | 2\*1280\*16\*16 | 1682.3 MiB, 1722648.2 KiB  (2\*1280\*16\*16)\*2/1024=1280 KiB  1721368.5+1280=1722648.2 KiB | **1682.3 MiB, 1722648.5 KiB** | | 30 | Transformer | 2\*1280\*16\*16 | 2\*1280\*16\*16  （覆盖，不申请内存） | 1682.3 MiB, 1722648.2 KiB | **1682.3 MiB, 1722648.5 KiB** | | 31 | ResnetBlock2D | 2\*1280\*16\*16  2\*640\*16\*16 | 2\*1280\*16\*16 | 1682.3 MiB, 1722648.2 KiB | **1682.3 MiB, 1722648.5 KiB** | | 32 | Transformer | 2\*1280\*16\*16 | 2\*1280\*16\*16  （覆盖，不申请内存） | 1682.3 MiB, 1722648.2 KiB | **1682.3 MiB, 1722648.5 KiB** | | 33 | Interpolate2D | 2\*1280\*16\*16 | 2\*1280\*32\*32 | **1687.3 MiB, 1727768.52KiB**  （2\*1280\*32\*32)\*2/1024=5120 KiB  1722648.2+5120=1727768.2 KiB | **1687.3 MiB, 1727768.5 KiB** | | 34 | CrossAttnUPBlock2D-2 | ResnetBlock2D | 2\*1280\*32\*32  2\*640\*32\*32 | 2\*640\*32\*32 | **1686.6 MiB, 1727128.2 KiB** | **1686.6 MiB, 1727128.5 KiB** | | 35 | Transformer | 2\*640\*32\*32 | 2\*640\*32\*32  （覆盖，不申请内存） | **1686.6 MiB, 1727128.2 KiB** | **1686.6 MiB, 1727128.5 KiB** | | 36 | ResnetBlock2D | 2\*640\*32\*32  2\*640\*32\*32 | 2\*640\*32\*32 | 1689.1 MiB, 1729688.2 KiB  (2\*640\*32\*32)\*2/1024=2560 KiB  1727128.2+2560=1729688.2 KiB | **1689.1 MiB, 1729688.5 KiB** | | 37 | Transformer | 2\*640\*32\*32 | 2\*640\*32\*32  （覆盖，不申请内存） | **1689.1 MiB, 1729688.2 KiB** | **1689.1 MiB, 1729688.5 KiB** | | 38 | ResnetBlock2D | 2\*320\*32\*32  2\*640\*32\*32 | 2\*640\*32\*32 | **1689.1 MiB, 1729688.2 KiB** | **1689.1 MiB, 1729688.5 KiB** | | 39 | Transformer | 2\*640\*32\*32 | 2\*640\*32\*32  （覆盖，不申请内存） | **1689.1 MiB, 1729688.2 KiB** | **1689.1 MiB, 1729688.5 KiB** | | 40 | Interpolate2D | 2\*640\*32\*32 | 2\*640\*64\*64 | **1699.1 MiB, 1739928.2 KiB**  (2\*640\*64\*64)\*2/1024=10240 KiB  1729688.2+10240=1739928.2 KiB | **1699.1 MiB, 1739928.5 KiB** | | 41 | CrossAttnUPBlock2D-3 | ResnetBlock2D | 2\*640\*64\*64  2\*320\*64\*64 | 2\*320\*64\*64 | **1687.9 MiB, 1728408.2 KiB** | **1687.9 MiB, 1728408.5 KiB** | | 42 | Transformer | 2\*320\*64\*64 | 2\*320\*64\*64  （覆盖，不申请内存） | **1687.9 MiB, 1728408.2 KiB** | **1687.9 MiB, 1728408.5 KiB** | | 43 | ResnetBlock2D | 2\*320\*64\*64  2\*320\*64\*64 | 2\*320\*64\*64 | **1692.9 MiB, 1733528.2 KiB**  (2\*320\*64\*64)\*2/1024=5120 KiB  1728408.2+5120=1733528.2 KiB | **1692.9 MiB, 1733528.5 KiB** | | 44 | Transformer | 2\*320\*64\*64 | 2\*320\*64\*64  （覆盖，不申请内存） | **1692.9 MiB, 1733528.5 KiB** | **1692.9 MiB, 1733528.5 KiB** | **3806.0 MiB** | **4426.0 MiB** | | 45 | ResnetBlock2D | 2\*320\*64\*64  2\*320\*64\*64 | 2\*320\*64\*64 | **1692.9 MiB, 1733528.2 KiB** | **1692.9 MiB, 1733528.5 KiB** | | 46 | Transformer | 2\*320\*64\*64 | 2\*320\*64\*64  （覆盖，不申请内存） | **1692.9 MiB, 1733528.2 KiB** | **1692.9 MiB, 1733528.5 KiB** | | 48 | Post\_process | Groupnorm | 2\*320\*64\*64 | 2\*320\*64\*64 | **1682.9 MiB, 1723608.2 KiB**  (2\*320\*64\*64)\*2/1024=5120 KiB  1718488.2+5120=1723608.2 KiB | **1682.9 MiB, 1723288.5 KiB** | | 49 | SiLU | 2\*320\*64\*64 | 2\*320\*64\*64 | **1687.9 MiB, 1728408.2 KiB**  (2\*320\*64\*64)\*2/1024=5120 KiB  1723288.2+5120=1728408.2 KiB | **1687.9 MiB, 1728408.5 KiB** | | 50 | Conv2D | 2\*320\*64\*64 | 2\*4\*64\*64 | **1688.0 MiB, 1728472.2 KiB**  (2\*4\*64\*64)\*2/1024=64 KiB  1728408.2+64=1728472.2 KiB | **1688.0 MiB, 1728472.5 KiB** | |