

--- Tiny example (no mask) ---

scores:

```
tensor([[0.7071, 0.0000, 1.4142],  
        [0.0000, 0.7071, 0.0000],  
        [1.4142, 0.0000, 2.8284]])
```

weights:

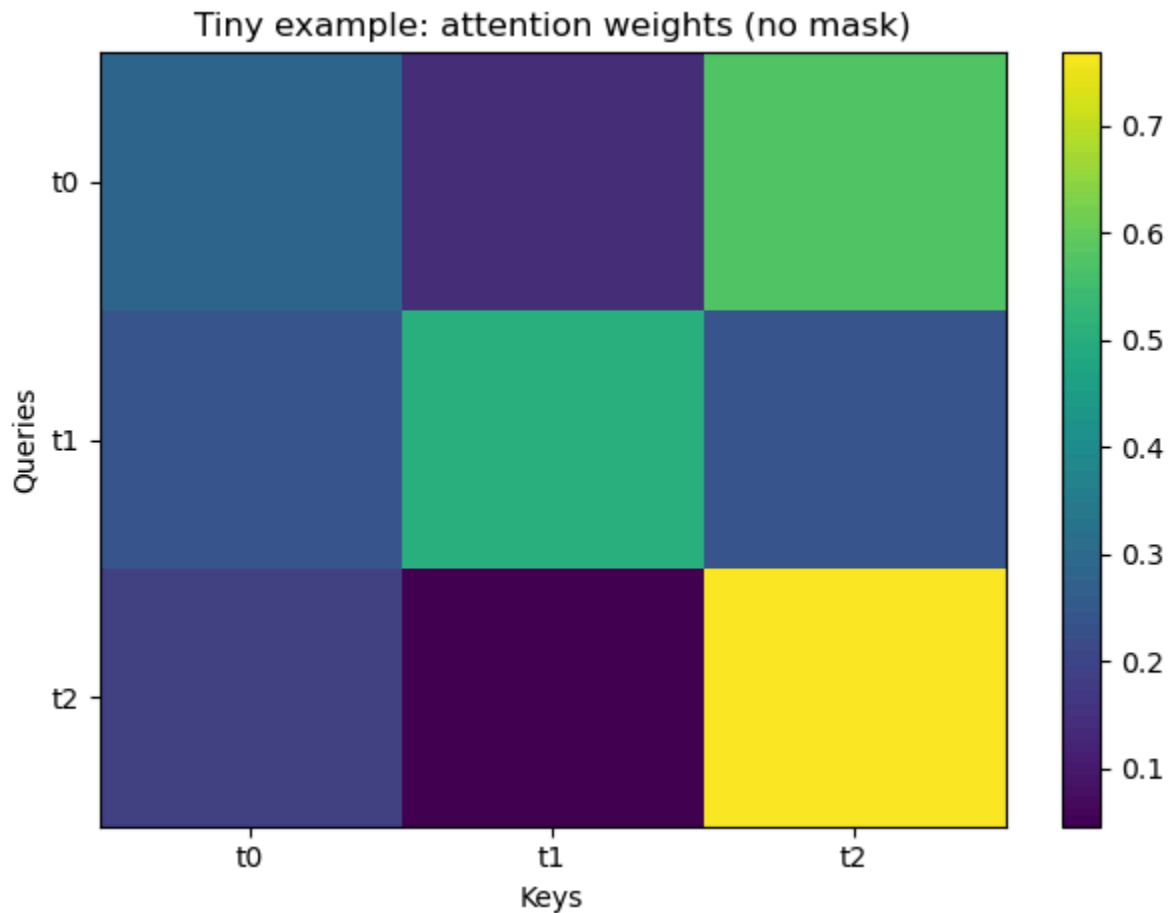
```
tensor([[0.2840, 0.1400, 0.5760],  
        [0.2483, 0.5035, 0.2483],  
        [0.1867, 0.0454, 0.7679]])
```

row sums:

```
tensor([1., 1., 1.])
```

out:

```
tensor([[8.5997, 7.1600],  
        [4.9651, 7.5174],  
        [9.5461, 8.1331]])
```



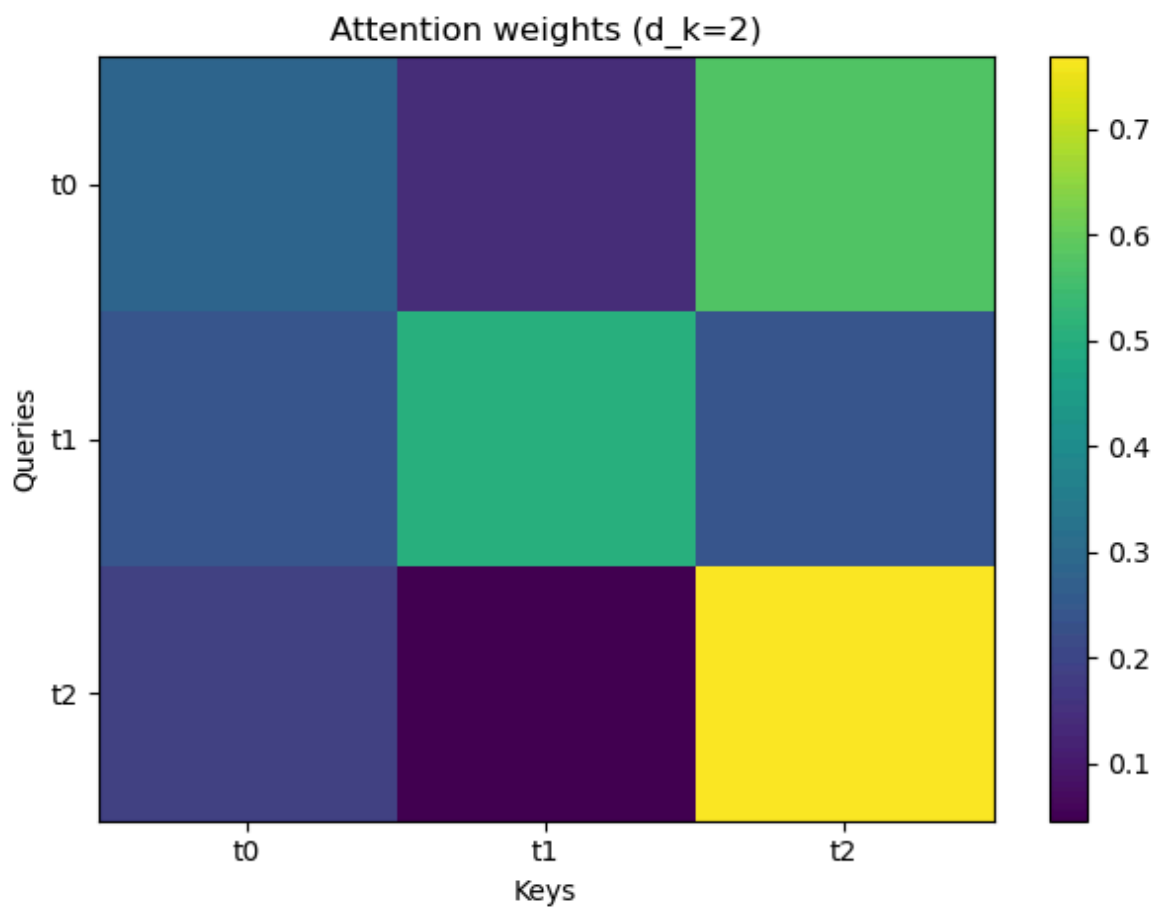
--- d_k experiment: d_k=2 ---

scores:

```
tensor([[0.7071, 0.0000, 1.4142],  
        [0.0000, 0.7071, 0.0000],  
        [1.4142, 0.0000, 2.8284]])
```

weights:

```
tensor([[0.2840, 0.1400, 0.5760],  
        [0.2483, 0.5035, 0.2483],  
        [0.1867, 0.0454, 0.7679]])
```



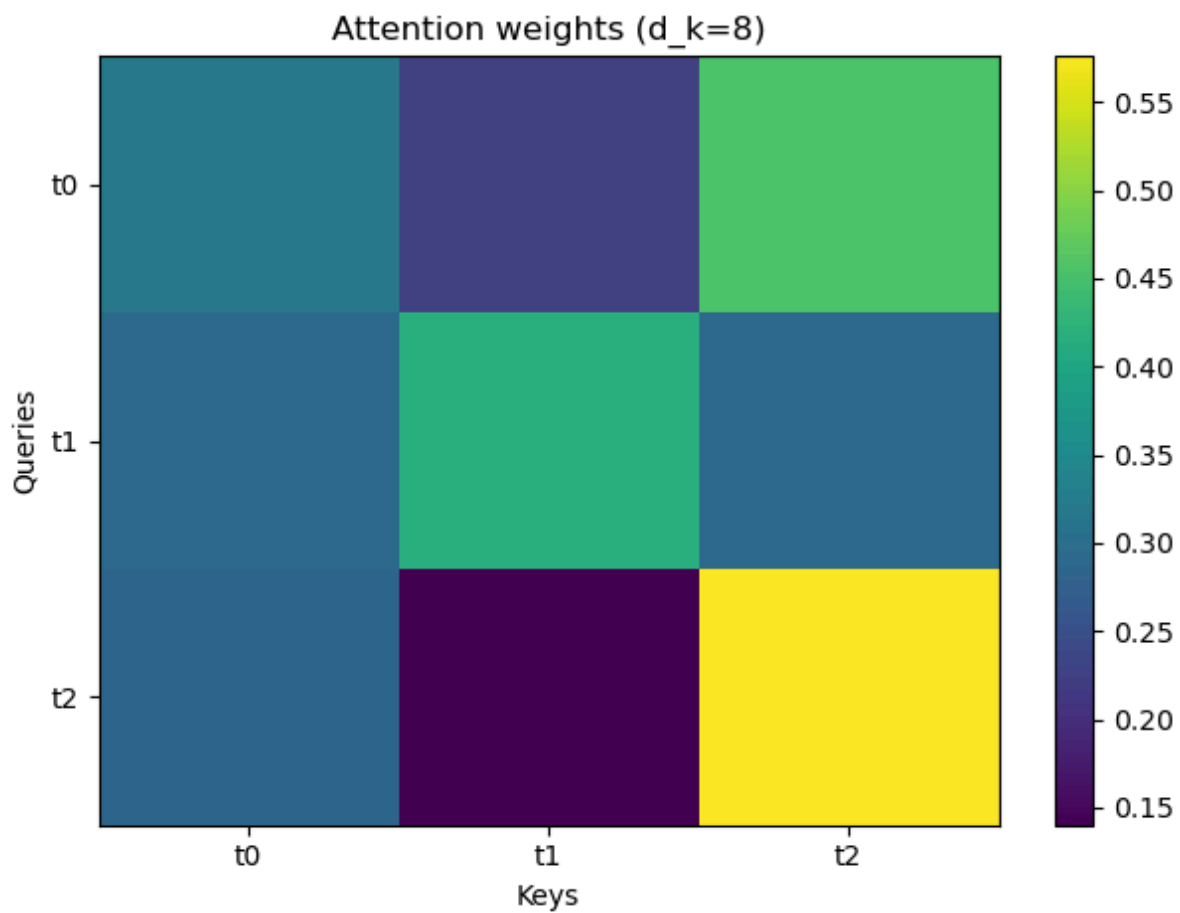
--- d_k experiment: d_k=8 ---

scores:

```
tensor([[0.3536, 0.0000, 0.7071],  
        [0.0000, 0.3536, 0.0000],  
        [0.7071, 0.0000, 1.4142]])
```

weights:

```
tensor([[0.3199, 0.2246, 0.4555],  
        [0.2920, 0.4159, 0.2920],  
        [0.2840, 0.1400, 0.5760]])
```



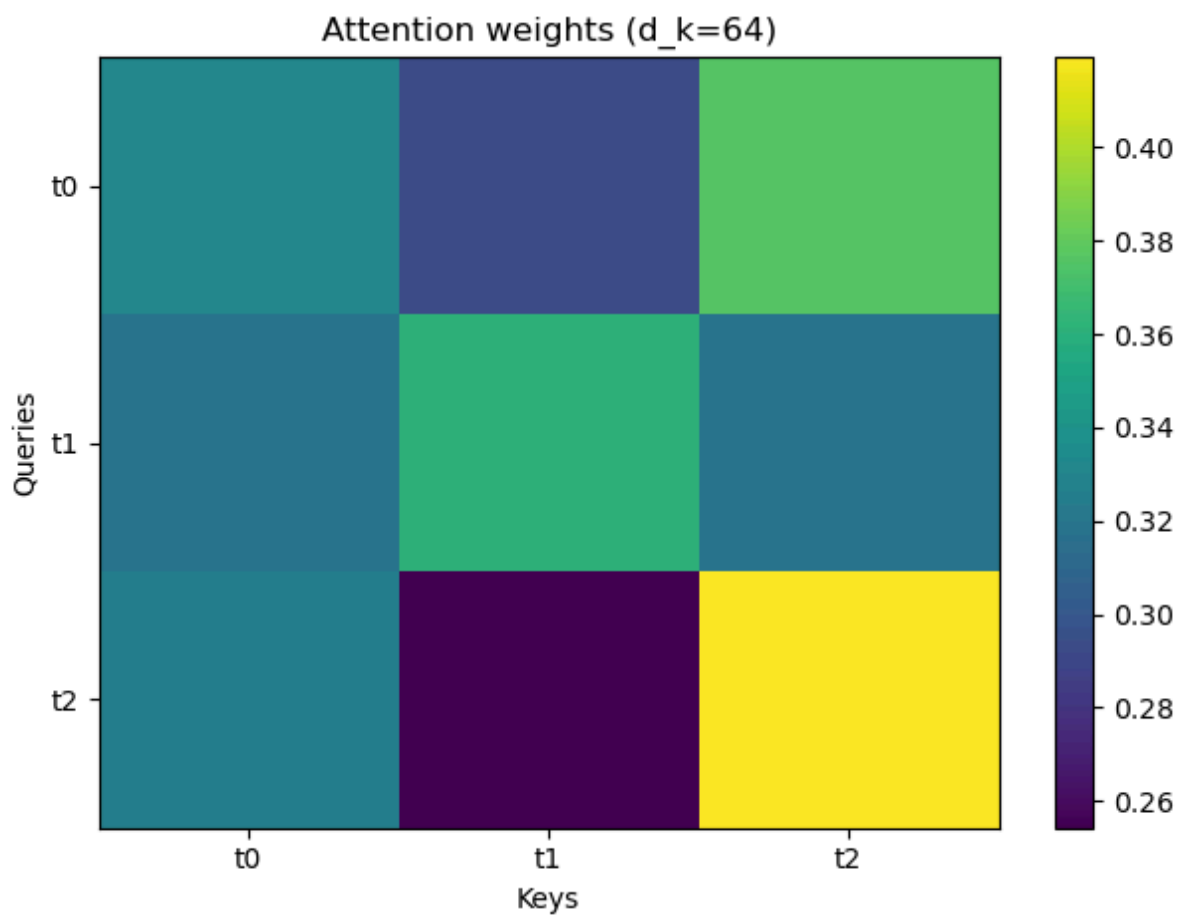
--- d_k experiment: d_k=64 ---

scores:

```
tensor([[0.1250, 0.0000, 0.2500],  
        [0.0000, 0.1250, 0.0000],  
        [0.2500, 0.0000, 0.5000]])
```

weights:

```
tensor([[0.3316, 0.2926, 0.3758],  
        [0.3192, 0.3617, 0.3192],  
        [0.3265, 0.2543, 0.4192]])
```



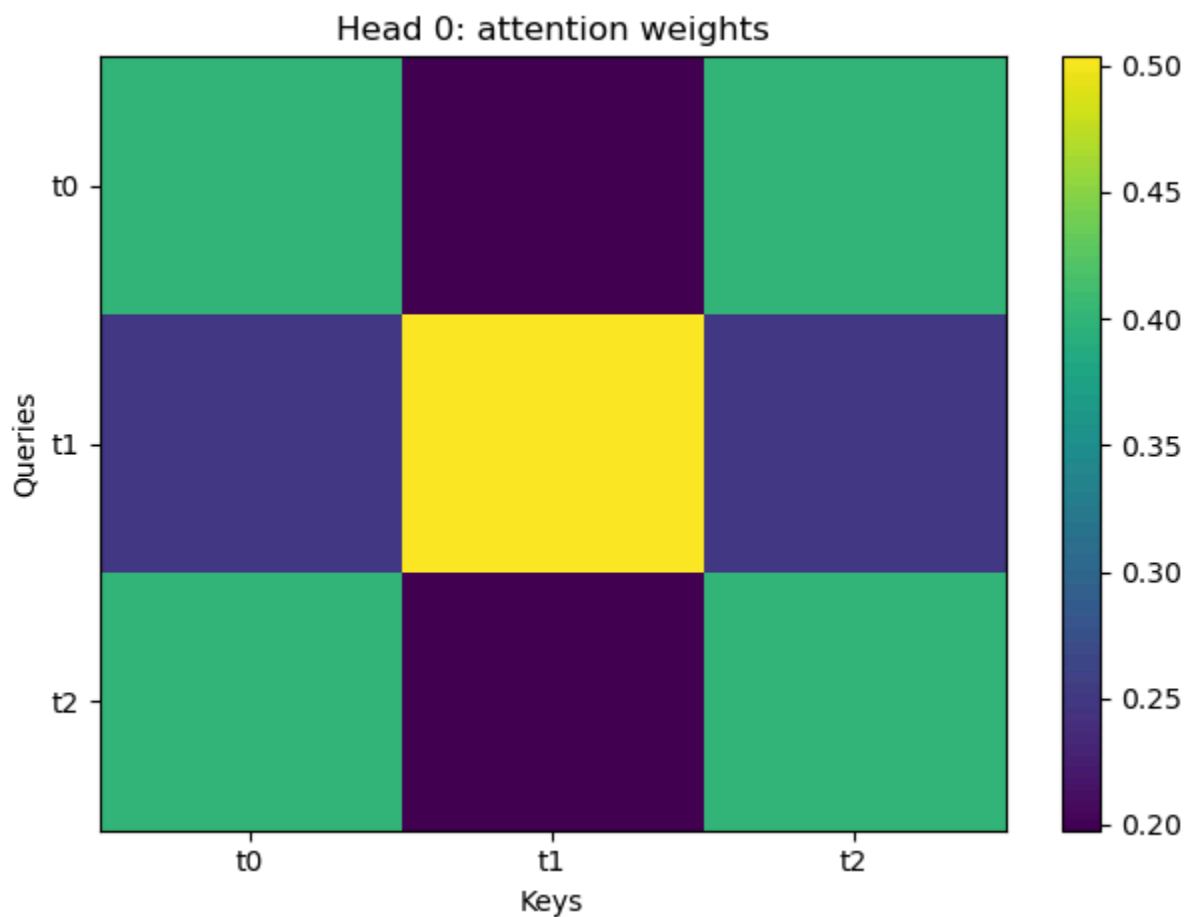
--- Multi-head self-attention (2 heads) ---

out (concat heads):

```
tensor([[0.8022, 0.1978, 0.4965, 0.5035],  
        [0.4965, 0.5035, 0.8022, 0.1978],  
        [0.8022, 0.1978, 0.8022, 0.1978]])
```

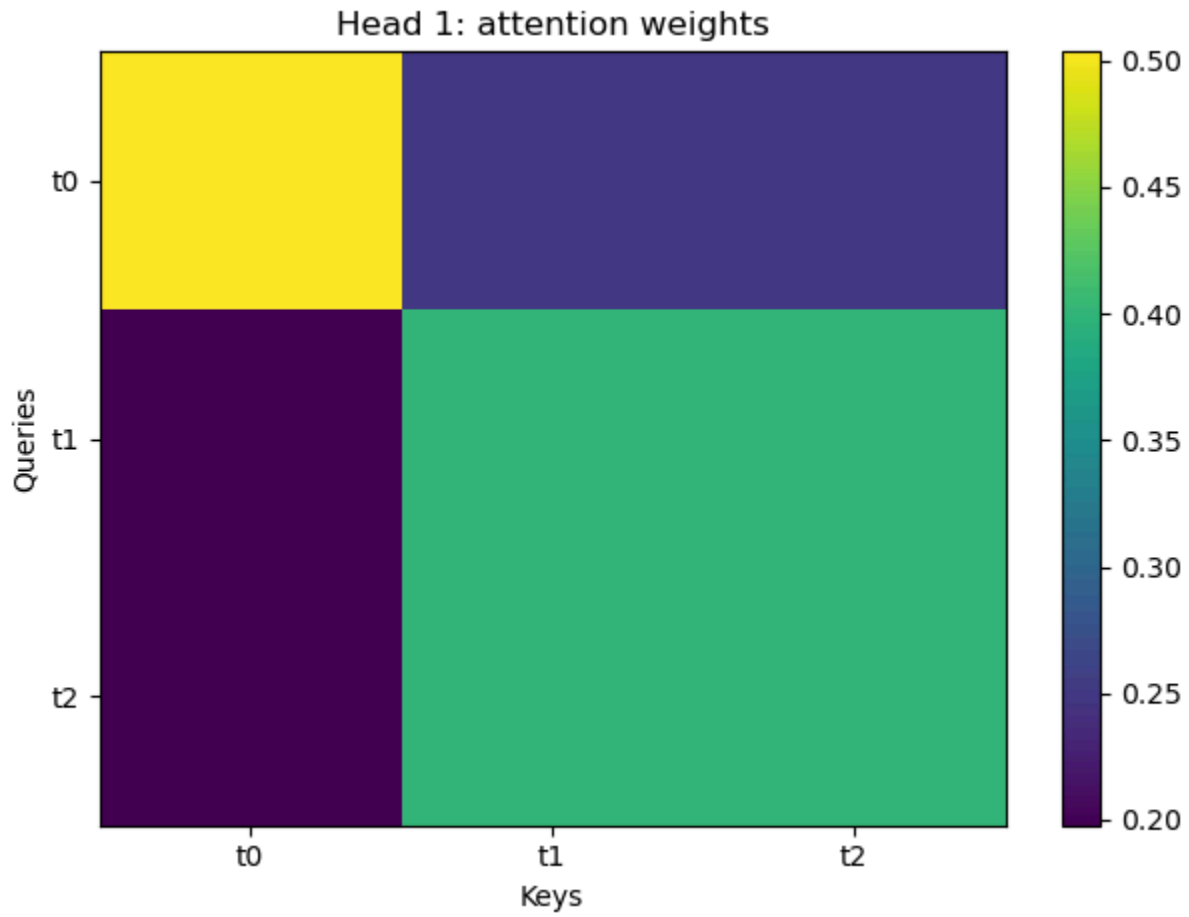
Head 0 weights:

```
tensor([[0.4011, 0.1978, 0.4011],  
        [0.2483, 0.5035, 0.2483],  
        [0.4011, 0.1978, 0.4011]])
```



Head 1 weights:

```
tensor([[0.5035, 0.2483, 0.2483],  
        [0.1978, 0.4011, 0.4011],  
        [0.1978, 0.4011, 0.4011]])
```



--- Self-attention (Q=K=V) ---

scores:

```
tensor([[0.7071, 0.0000, 0.7071],
        [0.0000, 0.7071, 0.0000],
        [0.7071, 0.0000, 0.7071]])
```

weights:

```
tensor([[0.4011, 0.1978, 0.4011],
        [0.2483, 0.5035, 0.2483],
        [0.4011, 0.1978, 0.4011]])
```

out:

```
tensor([[0.8022, 0.1978],
        [0.4965, 0.5035],
        [0.8022, 0.1978]])
```

