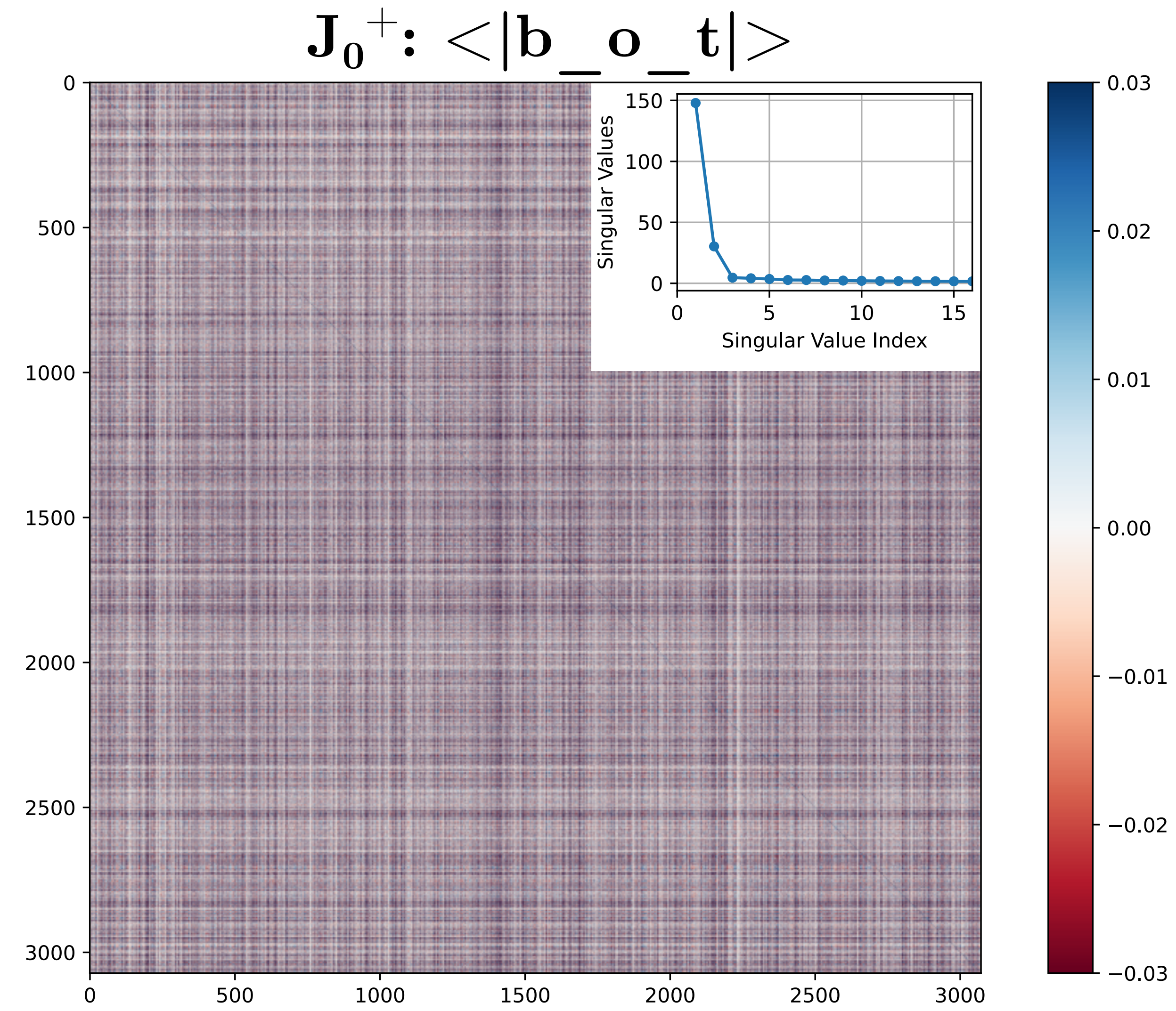
A

$t_0 = <|\text{begin\_of\_text}|>,\ x_0 = \text{embed}(t_0),\ y_{1T} = \text{The}$

$y_1(x) = \text{model}(x_0) = J_0^{+}(x)\cdot x_0$

One-token input: <|b_o_t|>
Predicted token: The

J_i^{+}: detached Jacobian matrix of the i-th token

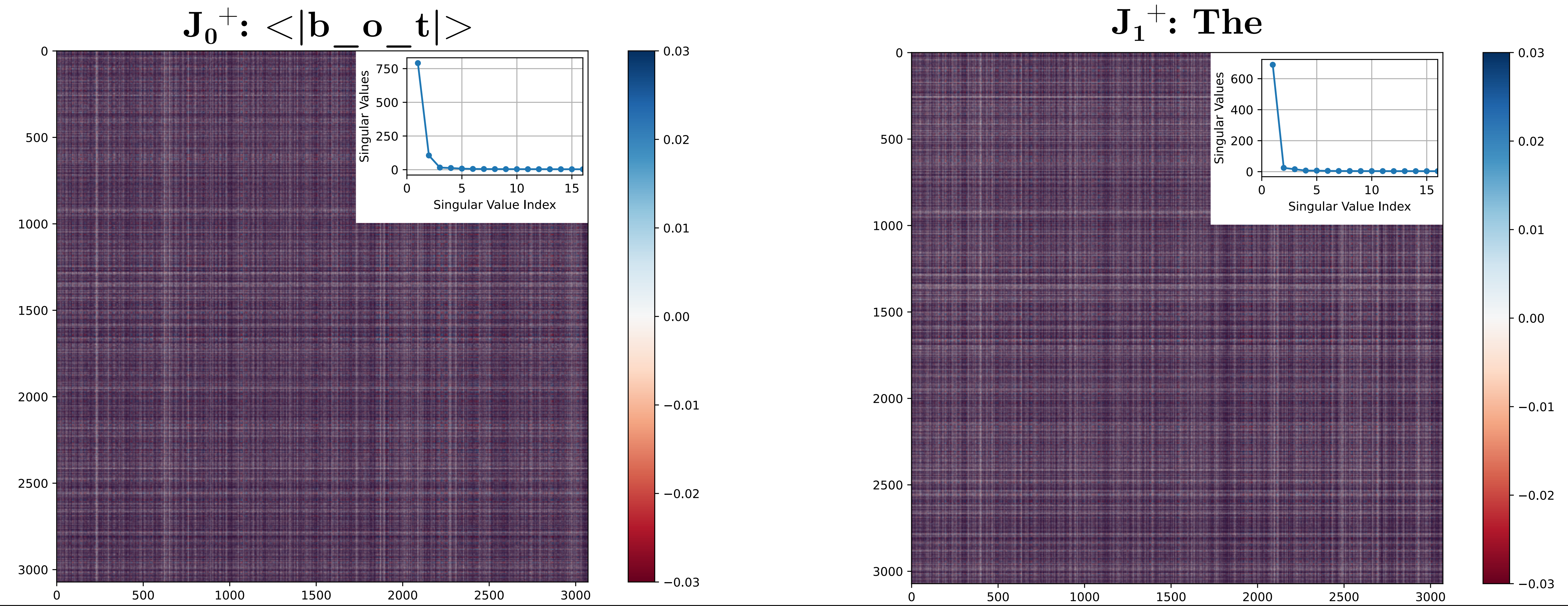


B

$t_0 = <|\text{begin\_of\_text}|>,\ t_1 = \text{The},\ (x_0, x_1) = \text{embed}(t_0, t_1),\ y_{2T} = \text{' '}$

$y_2(x) = \text{model}(x_0\ ,\ x_1)\ = J_0^{+}(x)\cdot x_0 + J_1^{+}(x)\cdot x_1$

Two-token input: :<|b_o_t|>, The
Predicted token: : ' '

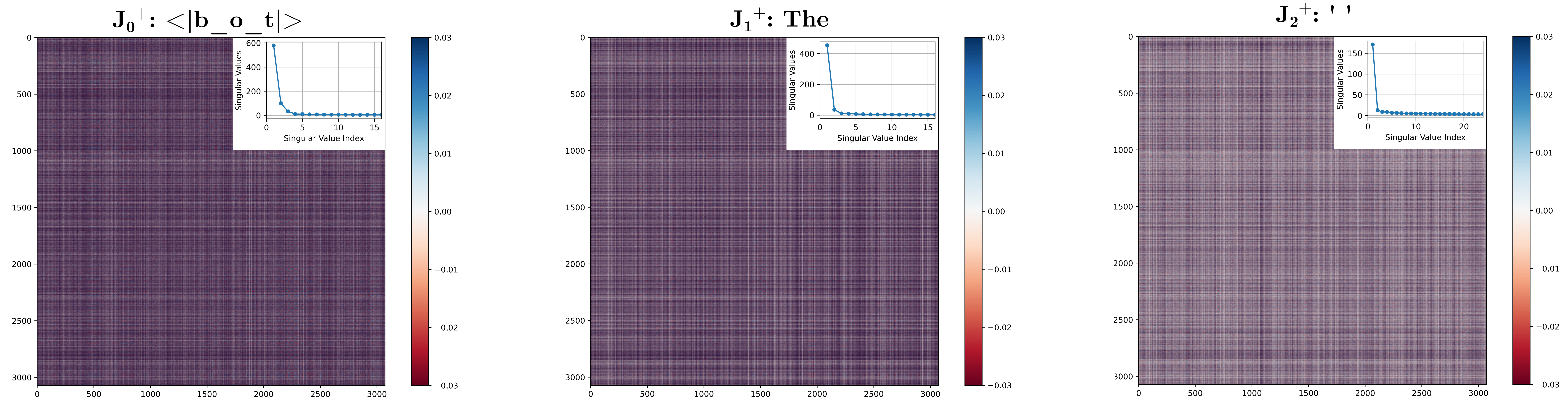


C

$t_0 = <|\text{begin\_of\_text}|>,\ t_1 = \text{The},\ t_2 = :\backslash n,\ (x_0, x_1, x_2) = \text{embed}(t_0, t_1, t_2),\ y_{3T} = 201$

$y_3(x) = \text{model}(x_0\ ,\ x_1\ ,\ x_2) = J_0^{+}(x)\cdot x_0 + J_1^{+}(x)\cdot x_1 + J_2^{+}(x)\cdot x_2$

Three-token input: <|b_o_t|>, The, ' '
Predicted token: 201


J_0^{+}: <|b_o_t|>
J_1^{+}: The
J_2^{+}: ' '
Singular Values
Singular Value Index
0
50
100
150
200
400
600
5
10
15
20
0
500
1000
1500
2000
2500
3000
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03