Generalization Error Bounds for Picard-Type Operator Learning in Nonlinear Parabolic PDEs

Koichi Taniguchi
Department of Mathematical and Systems Engineering, Faculty of Engineering
Shizuoka University
3-5-1 Johoku, Chuo-ku, Hamamatsu, 432-8561, Japan
taniguchi.koichi@shizuoka.ac.jp

Sho Sonoda
RIKEN AIP / CyberAgent
1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
sho.sonoda@riken.jp

Abstract

Operator learning for partial differential equations (PDEs) aims to learn solution operators on infinite-dimensional function spaces from finite-resolution data. In this setting, it is important for the learned model to be discretization-invariant, or resolution-robust, and to reflect PDE-specific structure. It is therefore natural to ask how such structure should be encoded in the model architecture, hypothesis class, or learning procedure. In this paper, we study operator learning for solution operators of nonlinear parabolic PDEs based on Duhamel–Picard iteration. We formulate Picard iteration as an abstract state-transition model and present a theoretical framework for Picard-type operator learning. We derive implementation-agnostic generalization error bounds that separate the implementation error from the estimation error associated with the abstract state-transition model induced by Picard iteration. A key consequence is that increasing the Picard depth reduces the Picard truncation error without causing an unbounded growth of the entropy-based estimation error. We also extend the analysis to long-time prediction by rolling out the same learned local model over successive time blocks. Finally, we illustrate the theory for nonlinear heat equations on the torus using a Picard-type Fourier neural operator as a concrete implementation.

1 Introduction

Operator learning is a deep learning framework for approximating maps between infinite-dimensional function spaces. It has become a central approach to building surrogate solvers for partial differential equations (PDEs) [Lu et al., 2021, Li et al., 2021, Kovachki et al., 2023, 2024b]. Once trained, such models can predict the solution for a new input without rerunning a classical numerical solver, which is useful in repeated-query tasks such as uncertainty quantification, design optimization, inverse problems, and control. Unlike standard high-dimensional regression, however, the object to be learned is an operator on function spaces. Even when the underlying inputs and outputs are naturally viewed as functions, training data are often accessed through finite-resolution representations, such as sensor values, grid values, or truncated coefficients. A theory of operator learning must therefore account both for statistical generalization and for robustness with respect to the choice of sensors or discretization.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

For PDE surrogate modeling there is an additional issue: the target is not an arbitrary continuous operator. It is generated by an equation and carries analytic structure, such as semigroup and Duhamel formulas, stability estimates, fixed-point iterations, and sometimes comparison or conservation principles. This raises a basic statistical question. If an operator learning model is aligned with the solver structure of the PDE, can increasing its depth improve approximation without paying a growing statistical price? In generic deep-learning bounds, depth can enlarge the hypothesis class and worsen capacity estimates. In solver-aligned operator learning, by contrast, additional depth may simply perform more steps of a stable iterative scheme.

This paper studies this question for nonlinear parabolic PDEs whose local solution operators are obtained by Duhamel–Picard iteration. The work is motivated by two complementary ideas. First, Furuya et al. [2025] showed that neural-operator layers can be aligned with Picard steps and used this alignment to obtain quantitative approximation results without the usual exponential growth of parametric complexity for general operators. Second, Sonoda et al. [2025, Section 6.2] identifies this Picard-aligned neural operator setting as a favorable instance of an implementation-agnostic state-transition framework: Picard iteration gives fast depth-driven approximation, while contractivity keeps the transition word-ball entropy controlled. The present paper turns that observation into a finite-sample generalization theory for Picard-type operator learning.

Our analysis separates three levels of the learning problem. The target concept is the PDE solution operator. The statistical hypothesis class is an abstract Picard class generated by admissible equation specifications and a finite number of Picard steps. A concrete neural operator is treated as an implementation of this abstract Picard class. This distinction is important: the main bounds are implementation-agnostic oracle bounds for predictors obtained by empirical risk minimization over the abstract Picard class and then realized by a concrete implementation map. Thus architecture-specific approximation or optimization issues do not disappear; they enter through an implementation error. The statistical term, however, is controlled by the Rademacher complexity of the Picard transition class rather than by the parameter complexity of a particular neural-operator architecture.

The resulting bounds decompose the prediction error into implementation error, Picard truncation error, statistical estimation error, and, in the finite-observation setting, reconstruction error from sensor data. Under the uniform contraction assumptions used throughout the paper, the Picard truncation error decays geometrically with the Picard depth ℓ, whereas the entropy-based Rademacher bound remains controlled independently of ℓ. This gives a concrete learning-theoretic explanation of when depth is useful in operator learning: depth is beneficial when it follows a stable solver structure, so that approximation improves while transition complexity saturates or grows slowly.

Our contributions are as follows. First, we formulate Picard-type operator learning for nonlinear parabolic PDEs in a setting where the underlying equation may be unknown but is assumed to belong to an admissible equation class. The goal is not to identify the equation itself, but to learn the corresponding solution operator using the Picard iteration structure as structural prior information. Second, we formulate this structure as an abstract state-transition model and connect PDE-specific quantities, such as contraction and stability constants, with statistical complexity through the Rademacher complexity of the abstract Picard hypothesis class. Third, we prove generalization error bounds that separate the implementation error, the Picard transition model error, the statistical estimation error, and, in the finite-observation setting, the reconstruction error from sensor data. Fourth, we show that increasing the Picard depth reduces the Picard transition model error without increasing the entropy-based Rademacher term, and we give a practical corollary describing how to balance Picard depth, sample size, and sensor resolution. Fifth, we extend the analysis to long-time prediction by rolling out the learned local model over time slabs and identify stability conditions under which the error remains controlled. Finally, we illustrate how the general theory can be used in a concrete setting by treating nonlinear heat equations on the torus with Fourier neural operators, where we derive explicit implementation error and generalization error bounds.

2 Related Work

In operator learning, various models have been proposed, such as DeepONet [Lu et al., 2021, Lanthaler et al., 2022], PCA-Net [Bhattacharya et al., 2021, Lanthaler, 2023], and neural operators [Li et al., 2020, 2021, Kovachki et al., 2021, Gupta et al., 2021, Bonev et al., 2023, Tripura and Chakraborty, 2023, Chen et al., 2024]. Their universal approximation properties have also been studied [Chen and

Chen, 1993, 1995, Kovachki et al., 2021, Lanthaler et al., 2022, Lanthaler, 2023, Lanthaler et al., 2025]. On the other hand, when one considers general operators, the curse of parametric complexity, namely the exponential growth of model complexity caused by infinite dimensionality, may appear [Lanthaler and Stuart, 2026]. To avoid this difficulty, it is important to use additional structure of the target operator. Indeed, several works show that, by restricting the target to solution operators of PDEs, one can avoid the curse of parametric complexity in certain problem classes [Kovachki et al., 2021, Chen et al., 2023, Lanthaler, 2023, Marcati and Schwab, 2023, Schwab et al., 2026, Furuya and Kratsios, 2024, Marcati and Schwab, 2024, Lanthaler and Stuart, 2026]. Generalization error and capacity analysis for operator learning have also been studied [Lanthaler et al., 2022, Gopalani et al., 2022, Kim and Kang, 2024, Lara Benitez et al., 2024, Kovachki et al., 2024a].

A work closely related to the present paper is Furuya et al. [2025]. They derived approximation error bounds for neural operators without the curse of parametric complexity by identifying forward propagation through the layers with the steps of Picard iteration. In this sense, their result provides an approximation-theoretic foundation for Picard-aligned neural operators. Our work is also related to approaches that incorporate fixed-point structures into learning models, such as deep equilibrium models [Bai et al., 2019, Marwah et al., 2023] and neural general operator networks based on Banach fixed point iterations [Feischl et al., 2025]. For long-time prediction, related experimental studies include Michałowska et al. [2023] and Lippe et al. [2023], which study error accumulation and long rollout behavior in neural PDE solvers.

This paper is based on the viewpoint that, in learning solution operators, the alignment between the internal structure of the solution operator and the model design is important. From this viewpoint, we analyze the Rademacher complexity of the abstract solution class generated by Picard iteration and derive generalization error bounds by using the implementation-agnostic framework of Sonoda et al. [2025]. This work provides a theoretical framework for understanding the generalization behavior of solver-aligned operator learning. See Appendix A for more details.

3 Problem Setting

In this section, we describe the operator learning setting addressed in this paper. Let $d \in \mathbb{N}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $T > 0$ and $Q_T := D \times (0, T)$. We consider the initial-value problem of nonlinear parabolic PDEs:

$$\begin{cases} \partial_t u + Au = F(u) & \text{in } Q_T, \\ u(0) = u_0 & \text{in } D, \end{cases} \tag{P}$$

where A is a linear operator and $F : \mathbb{R} \to \mathbb{R}$ is a nonlinear function satisfying $F(0) = 0$. Concrete examples are given in Section 7 and Appendix B.

In this paper, we allow the underlying equation to be unknown. Let $\mathcal{A}$ be a class of admissible linear operators, and let $\mathcal{F}$ be a class of admissible nonlinearities. We define the admissible equation class by $\Omega_{\mathrm{eq}} := \mathcal{A} \times \mathcal{F}$. An element $\omega = (A, F) \in \Omega_{\mathrm{eq}}$ is called an admissible equation specification. For each $\omega = (A, F)$, we denote by $\mathcal{G}_\omega$ the solution operator to (P). The unknown target operator $\mathcal{G}_*$ is assumed to be generated by some $\omega_* = (A_*, F_*) \in \Omega_{\mathrm{eq}}$, that is, $\mathcal{G}_* = \mathcal{G}_{\omega_*}$. The case where the equation is known corresponds to taking $\Omega_{\mathrm{eq}} = \{(A_*, F_*)\}$. The purpose of this paper is not to identify the equation specification ω_*. Rather, the goal is to learn the solution operator $\mathcal{G}_*$ from data (see also Appendix G). The equation class Ω_{eq} is used as structural prior information for defining the operator class and for deriving generalization error bounds.

3.1 Target Operator Class

By the Duhamel principle, the problem (P) is formally equivalent to the integral formulation

$$u(t) = S_A(t)u_0 + \int_0^t S_A(t-\tau)F(u(\tau))\, d\tau, \tag{P'}$$

where $S_A(t)$ is the semigroup generated by A. A solution to (P') is often called a mild solution to (P). In this paper, we take $U_{0,R} := \{u_0 \in C(D) : \|u_0\|_{L^\infty(D)} \leqslant R\}$ and $U_M := \{u \in C([0,T]; C(D))) : \|u\|_{L^\infty(Q_T)} \leqslant M\}$, with $R, M > 0$, as the initial-data space and the solution space, respectively. For

$\omega = (A, F) \in \Omega_{\mathrm{eq}}$ and $u_0 \in U_{0,R}$, define the Picard map T_{ω,u_0} by

$$T_{\omega,u_0}[v](t) := S_A(t)u_0 + \int_0^t S_A(t-\tau)F(v(\tau))\,d\tau. \tag{1}$$

In this paper, we impose the following assumption on Ω_{eq}.

Assumption 3.1. There exist $R, M, T, C_S > 0$ and $\delta \in (0,1)$ such that, for any $\omega = (A, F) \in \Omega_{\mathrm{eq}}$ and $u_0 \in U_{0,R}$, the conditions (a) and (b) hold:

(a) The map $\mathcal{T}_{\omega,u_0} : U_M \to U_M$ is δ-contractive, i.e.,
$$\|\mathcal{T}_{\omega,u_0}[u] - \mathcal{T}_{\omega,u_0}[v]\|_{L^\infty(Q_T)} \leqslant \delta \|u - v\|_{L^\infty(Q_T)} \tag{2}$$
for any $u, v \in U_M$.

(b) The map $\mathcal{T}_{\omega,u_0}$ is Lipschitz continuous with respect to u_0, i.e.,
$$\|\mathcal{T}_{\omega,u_0}[u] - \mathcal{T}_{\omega,v_0}[u]\|_{L^\infty(Q_T)} \leqslant C_S \|u_0 - v_0\|_{L^\infty(D)} \tag{3}$$
for any $u_0, v_0 \in U_{0,R}$ and $u \in U_M$.

Assumption 3.1 consists of standard conditions for applying Banach's fixed point theorem to establish local well-posedness for nonlinear PDEs [Henry, 1981, Pazy, 1983]. The condition (a) guarantees the existence and uniqueness of the solution as a fixed point, while (b), together with (a), yields Lipschitz continuous dependence on the initial data. The condition (b) depends only on the linear operator A, independently of F, and is ensured by the uniform bound $\sup_{A\in\mathcal{A}} \sup_{0<t\leqslant T} \|S_A(t)\|_{L^\infty(D)\to L^\infty(D)} \leqslant C_S$. Various examples of admissible equation classes satisfying this assumption, together with detailed verification of these conditions, are given in Appendix B.

By Banach's fixed point theorem, we have the following (see, e.g., [Banach, 1922, Zeidler, 1986]).

Proposition 3.2. *Suppose Assumption 3.1 holds. Then the following statements hold:*

(i) *For any $\omega = (A, F) \in \Omega_{\mathrm{eq}}$, there exists a unique solution operator $\mathcal{G}_\omega : U_{0,R} \to U_M$ to* (P') *such that*
$$\|\mathcal{G}_\omega(u_0) - \mathcal{G}_\omega(v_0)\|_{L^\infty(Q_T)} \leqslant \frac{C_S}{1-\delta} \|u_0 - v_0\|_{L^\infty(D)} \tag{4}$$
for any $u_0, v_0 \in U_{0,R}$.

(ii) *For each $\omega = (A, F) \in \Omega_{\mathrm{eq}}$ and $u_0 \in U_{0,R}$, the Picard iteration*
$$u^{(\ell)} := \mathcal{T}^{[\ell]}_{\omega,u_0}[0] = (\underbrace{\mathcal{T}_{\omega,u_0} \circ \cdots \circ \mathcal{T}_{\omega,u_0}}_{\ell \text{ times}})[0], \quad \ell = 1, 2, \cdots,$$
converges to the solution $u := \mathcal{G}_\omega(u_0)$ in $L^\infty(Q_T)$ as $\ell \to \infty$ with
$$\mathrm{d}(u^{(\ell)}, u) \leqslant \frac{\delta^\ell}{1-\delta} \mathrm{d}(u^{(1)}, 0), \quad \ell = 1, 2, \cdots. \tag{5}$$

Form the above, the target operator is the solution operator $\mathcal{G}_* : U_{0,R} \to U_M$ and it has the Picard iteration structure. In Section 7, we study semilinear heat equations on the torus as a concrete example.

3.2 Initial Data Distribution

We assume that the initial data u_0 are sampled from a probability measure Π supported on $U_0 := \{v \in H^{s_0}(D) : \|v\|_{H^{s_0}(D)} \leqslant R_0\}$ with $s_0 > d/2$. We note that the Sobolev embedding gives $H^{s_0}(D) \hookrightarrow C(\overline{D})$, since D is a bounded Lipschitz domain and $s_0 > d/2$. In particular, point evaluations are well-defined. Moreover, we choose $R_0 \leqslant C_{\mathrm{Sob}}^{-1} R$ so that $U_0 \subset U_{0,R}$, where $C_{\mathrm{Sob}} = C_{\mathrm{Sob}}(D, s_0) > 0$ is some constant in the Sobolev inequality.

Remark 3.3. A canonical example is obtained by conditioning a centered Gaussian process on the Sobolev ball U_0. If the covariance kernel K is chosen so that $u_0 \sim GP(0, K)$ has $H^{s_0}(D)$-valued sample paths, then the Gaussian law on the separable Hilbert space $H^{s_0}(D)$ has positive mass on every ball centered at the origin; hence $\mathbb{P}(\|u_0\|_{H^{s_0}(D)} \leqslant R_0) > 0$ [Giné and Nickl, 2015, Corollary 2.6.17]. For example, on bounded Lipschitz domains, Matérn kernels with smoothness parameter ν yield $H^s(D)$-sample paths for every $s < \nu$, while squared-exponential kernels allow arbitrary finite Sobolev smoothness; see, e.g., Henderson [2024].

3.3 Population and Empirical Risks

Ideal risks. Let $\Gamma : U_0 \to U_M$ be a candidate operator model for the target solution operator $\mathcal{G}_* : U_0 \to U_M$ associated with (P'). Let μ be a probability measure on Q_T describing the distribution of output query points. We define the ideal population risk by

$$L[\Gamma] := \mathbb{E}_{u_0 \sim \Pi, z \sim \mu} \left[|\Gamma(u_0)(z) - \mathcal{G}_*(u_0)(z)|^2 \right]. \tag{6}$$

The ideal risk corresponds to the full-information setting, in which the model Γ is evaluated as an operator acting on the whole initial function u_0, rather than on finitely many observations of u_0. This setting is natural for theoretical or synthetic datasets where u_0 is explicitly generated and the corresponding solution values can be evaluated.

Given i.i.d. trajectory blocks $(u_{0,i}, z_{i,1}, \ldots, z_{i,q}) \sim \Pi \otimes \mu^{\otimes q}$, the corresponding empirical risk is defined by

$$\widehat{L}_{n,q}[\Gamma] := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{q} \sum_{j=1}^{q} \ell_{\Lambda_{\mathrm{clip}}} \left(\Gamma(u_{0,i})(z_{i,j}), \mathcal{G}_*(u_{0,i})(z_{i,j}) \right).$$

Here, the clipped squared loss is defined by $\ell_{\Lambda_{\mathrm{clip}}}(a, y) := \min\{|a - y|^2, \Lambda_{\mathrm{clip}}\}$ with $\Lambda_{\mathrm{clip}} \geqslant 4M^2$. For $a, y \in [-M, M]$, we have $\ell_{\Lambda_{\mathrm{clip}}}(a, y) = |a - y|^2$. Thus, when predictions and target values lie in $[-M, M]$, the empirical and population risks use the same squared loss.

Risks in finite observations. In many practical settings, the input is available only through a finite-resolution representation, such as sensor or grid values. We model this by the observation map. Given sensor locations $\Xi_m := \{\xi_1, \ldots, \xi_m\} \subset D$, we define the observation map $P_m : H^{s_0}(D) \to \mathbb{R}^m$ by

$$P_m u_0 := (u_0(\xi_1), \cdots, u_0(\xi_m)) \in \mathbb{R}^m. \tag{7}$$

For a finite-observation model $\Gamma_m : \mathbb{R}^m \times Q_T \to \mathbb{R}$, the finite-observation population risk is defined by

$$L_m[\Gamma_m] := \mathbb{E}_{u_0 \sim \Pi, z \sim \mu} \left[|\Gamma_m(P_m u_0, z) - \mathcal{G}_*(u_0)(z)|^2 \right], \tag{8}$$

and the empirical risk is defined by

$$\widehat{L}_{m,n,q}[\Gamma_m] := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{q} \sum_{j=1}^{q} \ell_{\Lambda_{\mathrm{clip}}} \left(\Gamma_m(P_m u_{0,i}, z_{i,j}), \mathcal{G}_*(u_{0,i})(z_{i,j}) \right). \tag{9}$$

We also lift Γ_m to an operator model $\Gamma_m^{\uparrow} : U_0 \to U_M$ by $\Gamma_m^{\uparrow}(u_0)(z) := \Gamma_m(P_m u_0, z)$. Thus, even though the model only uses finite sensor values, its output is compared with the full solution value $\mathcal{G}_*(u_0)(z)$. In this sense, the target of evaluation remains the full solution operator. Note that $L[\Gamma_m^{\uparrow}] = L_m[\Gamma_m]$. This identity identifies the finite-observation risk with the restriction of the continuum-level risk to lifted, sensor-dependent models.

4 Picard Iteration as Abstract State-Transition Model

In this section, we rewrite the Picard iteration as an abstract state-transition model so that it fits the framework of Sonoda et al. [2025]. We also derive bounds for the Picard truncation error. The abstract Picard class introduced here will serve as the main hypothesis class in the generalization analysis in Section 5, where the estimation error is controlled by the Picard truncation error, viewed as an oracle term, together with the Rademacher complexity of abstract Picard class. The details and proofs for this section are given in Appendix C.

4.1 The ℓ-step Picard Transition Model

Define the state space $X := U_{0,R} \times Q_T \times U_M$. For $(v, z, u) \in X$, where $z = (x, t) \in Q_T$, define $f_\omega(v, z, u) := (v, z, \mathcal{T}_{\omega,v}[u])$, $J(v, z) := (v, z, 0)$, and $h(v, z, u) := u(z)$, where f_ω is the transition, J is the initial embedding, and h is the readout. Then the ℓ-step Picard transition model is

$$b_{\omega,\ell}(v, z) := h \circ f_\omega^{[\ell]} \circ J(v, z) = (\mathcal{T}_{\omega,v}^{[\ell]}[0])(z).$$

We define the abstract Picard class by $\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}} := \{b_{\omega,\ell} : \omega \in \Omega_{\mathrm{eq}}\}$.

Lemma 4.1 (Picard truncation error)**.** *Suppose that Assumption 3.1 holds. Then, for any $\omega \in \Omega_{\mathrm{eq}}$ and $\ell \in \mathbb{N}$, we have*

$$\sup_{v \in U_{0,R}} \|b_{\omega,\ell}(v,\cdot) - \mathcal{G}_\omega[v]\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta}.$$

Consequently, $\inf_{b \in \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} L[b] \leqslant M^2\delta^{2\ell}/(1-\delta)^2$.

4.2 The ℓ-step Finite-Observation Picard Transition Model

We consider the case where the learner observes only $P_m u_0$. Let $\mathcal{E}_m \subset \{E : \mathbb{R}^m \to U_0\}$ be a class of reconstruction maps. For $E \in \mathcal{E}_m$, define the state space $X_m := \mathbb{R}^m \times Q_T \times U_M$. For $(a,z,u) \in X_m$, define $f_{\omega,E}(a,z,u) := (a,z,\mathcal{T}_{\omega,E(a)}[u])$, $J_m(a,z) := (a,z,0)$, and $h_m(a,z,u) := u(z)$. Hence, the ℓ-step finite-observation Picard transition model is

$$b_{\omega,E,\ell}(a,z) := h_m \circ f_{\omega,E}^{[\ell]} \circ J_m(a,z) = (\mathcal{T}_{\omega,E(a)}^{[\ell]}[0])(z).$$

We define the abstract finite-observation Picard class by $\mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}} := \{b_{\omega,E,\ell} : \omega \in \Omega_{\mathrm{eq}}, E \in \mathcal{E}_m\}$. The reconstruction error is defined by $\varepsilon_{\mathrm{rec}}^2(m) := \inf_{E \in \mathcal{E}_m} \mathbb{E}_{u_0 \sim \Pi}[\|E(P_m u_0) - u_0\|_{L^\infty(D)}^2]$.

Lemma 4.2 (Picard truncation error)**.** *Suppose that Assumption 3.1 holds. Then, for any $\omega \in \Omega_{\mathrm{eq}}$, $E \in \mathcal{E}_m$, $u_0 \in U_0$ and $\ell \in \mathbb{N}$, we have*

$$\|b_{\omega,E,\ell}(P_m u_0,\cdot) - \mathcal{G}_\omega[u_0]\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta} + \frac{C_S}{1-\delta}\|E(P_m u_0) - u_0\|_{L^\infty(D)}.$$

Consequently, $\inf_{b \in \mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}}} L_m(b) \leqslant 2M^2\delta^{2\ell}/(1-\delta)^2 + 2C_S^2\varepsilon_{\mathrm{rec}}^2(m)/(1-\delta)^2$.

5 Generalization Error Bounds

Let $\mathfrak{N}_\ell \subset \{\Gamma : U_0 \to U_M\}$ and $\mathfrak{N}_{m,\ell} \subset \{\Gamma_m : \mathbb{R}^m \times Q_T \to \mathbb{R}\}$, and let $\mathrm{embed} : \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}} \to \mathfrak{N}_\ell$ and $\mathrm{embed}_m : \mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}} \to \mathfrak{N}_{m,\ell}$ be implementation embeddings from the abstract Picard classes into the corresponding implementation classes, respectively. Define the implementation errors by $\varepsilon_{\mathrm{imp}} := \sup_{b \in \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} |L[\mathrm{embed}\, b] - L[b]|$ and $\varepsilon_{\mathrm{imp}}^m := \sup_{b \in \mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}}} |L_m[\mathrm{embed}_m\, b] - L_m[b]|$.

Let $\mathcal{X}$ be an input space and let $\mathcal{B} \subset \{b : \mathcal{X} \times Q_T \to \mathbb{R}\}$. Let $\nu_{\mathcal{X}}$ be the distribution of the input variable on $\mathcal{X}$. For a trajectory-level sample $\mathcal{S}_{n,q}^{\mathcal{X}} = \{(a_i, \{z_{i,j}\}_{j=1}^q)\}_{i=1}^n$ with independent blocks $(a_i, z_{i,1}, \ldots, z_{i,q}) \sim \nu_{\mathcal{X}} \otimes \mu^{\otimes q}$, define the pointwise empirical Rademacher complexity by

$$\widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{\mathcal{X}}}^{\mathrm{pt}}(\mathcal{B}) := \mathbb{E}_\sigma\left[\sup_{b \in \mathcal{B}} \frac{1}{nq} \sum_{i=1}^n \sum_{j=1}^q \sigma_{i,j} b(a_i, z_{i,j})\right],$$

where $\{\sigma_{i,j} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant q\}$ are independent Rademacher variables. In the bounds below we use the trajectory-normalized version

$$\widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{\mathcal{X}}}(\mathcal{B}) := \sqrt{q}\, \widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{\mathcal{X}}}^{\mathrm{pt}}(\mathcal{B}) = \mathbb{E}_\sigma\left[\sup_{b \in \mathcal{B}} \frac{1}{n\sqrt{q}} \sum_{i=1}^n \sum_{j=1}^q \sigma_{i,j} b(a_i, z_{i,j})\right].$$

We use the following shorthand: $\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} := \widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{U_0}}(\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}})$ and $\widehat{\mathfrak{R}}_m^{\Omega_{\mathrm{eq}}} := \widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{P_m(U_0)}}(\mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}})$, with $\nu_{U_0} = \Pi$ and $\nu_{P_m(U_0)} = (P_m)_\# \Pi$.

The following theorem is an implementation-agnostic oracle bound in the sense of Sonoda et al. [2025]. The empirical minimization is performed over the abstract Picard class, and the selected abstract predictor is then realized by a concrete implementation map. Thus the bound does not claim that every parameter-space ERM over a neural-operator class enjoys the same guarantee; any architecture-specific approximation or optimization gap is represented through the implementation error.

Theorem 5.1 (Generalization error bounds)**.** *Suppose that Assumption 3.1 holds and that all predictors are clipped to $[-M, M]$. Then there exists a universal constant $C > 0$ such that, for every $0 < \zeta < 1$, the following statements hold with probability at least $1 - \zeta$:*

(i) *Let* $\widehat{b}_\ell \in \operatorname{argmin}_{b\in\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} \widehat{L}_{n,q}[b]$ *and set* $\widehat{\Gamma}_\ell := \mathrm{embed}(\widehat{b}_\ell) \in \mathfrak{N}_\ell$. *Then*

$$L[\widehat{\Gamma}_\ell] \leqslant \varepsilon_{\mathrm{imp}} + \frac{M^2\delta^{2\ell}}{(1-\delta)^2} + CM\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}.$$

(ii) *Let* $\widehat{b}_{m,\ell} \in \operatorname{argmin}_{b\in\mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}}} \widehat{L}_{m,n,q}[b]$ *and set* $\widehat{\Gamma}_{m,\ell} := \mathrm{embed}_m(\widehat{b}_{m,\ell}) \in \mathfrak{N}_{m,\ell}$. *Then*

$$L_m[\widehat{\Gamma}_{m,\ell}] \leqslant \varepsilon_{\mathrm{imp}}^m + \frac{2M^2\delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2}{(1-\delta)^2}\varepsilon_{\mathrm{rec}}^2(m) + CM\widehat{\mathfrak{R}}_m^{\Omega_{\mathrm{eq}}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}.$$

The proof is based on viewing the Picard class defined in Section 4 as an abstract state-transition model and applying the bias-variance decomposition in the framework of Sonoda et al. [2025, Theorem 1]. The empirical-process step is performed over the n independent trajectory blocks. Since the clipped squared loss is CM-Lipschitz in the scalar prediction on $[-M, M]$, the block loss $r \mapsto q^{-1}\sum_{j=1}^q \ell_{\Lambda_{\mathrm{clip}}}(r_j, y_j)$ is $CM/\sqrt{q}$-Lipschitz with respect to the Euclidean norm on $\mathbb{R}^q$; Maurer's vector contraction inequality [Maurer, 2016] then yields the terms $CM\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}}$ and $CM\widehat{\mathfrak{R}}_m^{\Omega_{\mathrm{eq}}}$. The concentration term is also at the trajectory-block level and therefore scales as $n^{-1/2}$. The model error is estimated by using Lemmas 4.1 and 4.2. The details of proof is given in Appendix D.

Remark 5.2. Theorem 5.1 applies to any admissible equation class Ω_{eq} satisfying Assumption 3.1. Therefore, as illustrated in Appendix B, it provides a general theoretical framework for Picard-type operator learning across a broad range of nonlinear parabolic PDEs (see Proposition B.1 and examples in Appendix B). In particular, the framework does not require the underlying equation specification to be identified; it only requires that the target equation belong to an admissible class satisfying the uniform Picard contraction condition.

A more explicit form. The Rademacher terms in Theorem 5.1 can be bounded by metric entropy with respect to the empirical pointwise metric

$$d_{\mathcal{S},2}(b,b') := \left(\frac{1}{nq}\sum_{i=1}^n\sum_{j=1}^q |b(a_i,z_{i,j}) - b'(a_i,z_{i,j})|^2\right)^{1/2}.$$

Dudley's entropy integral gives $\widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{\mathcal{X}}}(\mathcal{B}) \lesssim n^{-1/2}\int_0^{2M}\sqrt{\log N(\mathcal{B}, d_{\mathcal{S},2}, \varepsilon)}\,d\varepsilon$. Let $d_\Omega(\omega,\omega') := \sup_{u_0\in U_{0,R}}\sup_{u\in U_M}\|\mathcal{T}_{\omega,u_0}[u] - \mathcal{T}_{\omega',u_0}[u]\|_{L^\infty(Q_T)}$. If $\log N(\Omega_{\mathrm{eq}}, d_\Omega, \varepsilon) \leqslant \mathcal{H}_\Omega(\varepsilon)$, then, by applying the Rademacher-complexity estimate of Sonoda et al. [2025] to the abstract Picard transition class and using the contraction property in Assumption 3.1, the full-information Rademacher term satisfies

$$\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} \lesssim \frac{1}{\sqrt{n}}\int_0^{2M}\sqrt{\mathcal{H}_\Omega(c(1-\delta)\varepsilon)}\,d\varepsilon,$$

where $c > 0$ is a universal constant. In particular, this upper bound is independent of the Picard depth ℓ. For the finite-observation class, one obtains similarly $\widehat{\mathfrak{R}}_m^{\Omega_{\mathrm{eq}}} \lesssim \mathcal{I}_{\Omega,m}/\sqrt{n}$, where $\mathcal{I}_{\Omega,m}$ is the corresponding entropy integral involving both $(\Omega_{\mathrm{eq}}, d_\Omega)$ and the reconstruction class $\mathcal{E}_m$. This bound is also independent of ℓ. Thus increasing ℓ reduces the Picard truncation error while the entropy-based statistical bound does not increase with ℓ.

Moreover, if $\varepsilon_{\mathrm{rec}}^2(m) \leqslant C_{\mathrm{rec}}^2 m^{-2\beta}$, then

$$L_m[\widehat{\Gamma}_{m,\ell}] \leqslant \varepsilon_{\mathrm{imp}}^m + \frac{2M^2\delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2C_{\mathrm{rec}}^2}{(1-\delta)^2}m^{-2\beta} + \frac{CM}{\sqrt{n}}\mathcal{I}_{\Omega,m} + CM^2\sqrt{\frac{\log(1/\zeta)}{n}}.$$

For quasi-uniform sensors and stable reconstruction on $U_0 \subset H^{s_0}(D)$, $s_0 > d/2$, one typically has $\beta = (s_0 - d/2)/d$.

Corollary 5.3 (Practical depth and sensor scaling)**.** *Assume the setting of Theorem 5.1. Suppose that the full-information entropy integral is finite,* $\mathcal{I}_\Omega := \int_0^{2M}\sqrt{\mathcal{H}_\Omega(c(1-\delta)\varepsilon)}\,d\varepsilon < \infty$, *and that the finite-observation entropy integral satisfies* $\mathcal{I}_{\Omega,m} \leqslant C_I m^\alpha$ *for some* $C_I > 0$ *and* $\alpha \geqslant 0$. *Suppose also*

that $\varepsilon_{\mathrm{rec}}^2(m) \leqslant C_{\mathrm{rec}}^2 m^{-2\beta}$ *for some* $\beta > 0$*. Let* $\ell_n := \lceil \log n/(4|\log \delta|) \rceil$*. Then the full-information bound with Picard depth* ℓ_n *takes the form*

$$L[\widehat{\Gamma}_{\ell_n}] \leqslant \varepsilon_{\mathrm{imp}} + C\left(\frac{M^2}{\sqrt{n}} + \frac{M\mathcal{I}_\Omega}{\sqrt{n}} + M^2\sqrt{\frac{\log(2/\zeta)}{n}}\right).$$

Moreover, if $m_n \asymp n^{1/(2(2\beta+\alpha))}$*, then the finite-observation bound with* $m = m_n$ *and* $\ell = \ell_n$ *takes the form*

$$L_{m_n}[\widehat{\Gamma}_{m_n,\ell_n}] \leqslant \varepsilon_{\mathrm{imp}}^{m_n} + C\left(n^{-\beta/(2\beta+\alpha)} + M^2\sqrt{\frac{\log(2/\zeta)}{n}}\right),$$

where the constant C *may depend on* $M, C_S, \delta, C_{\mathrm{rec}}, C_I$*, but not on* n*. In the quasi-uniform reconstruction example,* $\beta = (s_0 - d/2)/d$*.*

The details of proof is given in Appendix D. Corollary 5.3 gives the design rule suggested by the main theorem. Since the Picard truncation term is geometric, logarithmic depth $\ell_n = O(\log n)$ is enough to push it down to the statistical scale $n^{-1/2}$. The finite-observation case adds the usual trade-off between sensor resolution and statistical complexity: increasing m reduces reconstruction error but may enlarge the reconstruction-class entropy.

6 Long-Time Prediction by Local Rollout

In this section, we study long-time prediction by iterating the same learned local model $\widehat{\Gamma}_\ell$ over successive time blocks. Throughout this section, we assume that $\widehat{\Gamma}_\ell$, initially defined on U_0, extends to a map $\widehat{\Gamma}_\ell : U_{0,R} \to U_M$. We use the same notation for this extension. The resulting error bound depends on the rollout-relevant local error (10), which measures the uniform approximation error on the states visited by the rollout. Control of this error is an additional requirement and does not follow directly from the population risk bound in Theorem 5.1. Since the finite-observation case can be treated similarly, we discuss only the full-information case here.

Define $\psi : U_M \to L^\infty(D)$ by $\psi(u) := u(\cdot, T)$, and assume that $\|\psi(u) - \psi(v)\|_{L^\infty(D)} \leqslant L_\psi \|u - v\|_{L^\infty(Q_T)}$. Note that $L_\psi \leqslant 1$. Let $\Phi_* := \psi \circ \mathcal{G}_*$. For $\kappa \in \mathbb{N}$, define $U_0^{[\kappa]} := \{u_0 \in U_0 : \Phi_*^j(u_0) \in U_{0,R} \text{ for } j = 0, \ldots, \kappa\}$. This condition ensures that the same local solution theory can be applied on each time block. We assume $\Pi(U_0^{[\kappa]}) > 0$ and write $\Pi_\kappa := \Pi(\cdot \mid U_0^{[\kappa]})$.

For $u_0 \in U_0^{[\kappa]}$, define the exact states by $v_0 = u_0$ and $v_{j+1} = \Phi_*(v_j)$. The exact solution on the j-th block is $\mathcal{G}_*[v_j]$ after the time shift $t = jT + s$, $s \in [0, T]$. Next, define the approximate transition by $\widehat{\Phi}_\ell := \mathcal{P}_R \circ \psi \circ \widehat{\Gamma}_\ell$, where $\mathcal{P}_R : C(D) \to U_{0,R}$ is the pointwise clipping map. The approximate states are $\widehat{v}_0 = u_0$ and $\widehat{v}_{j+1} = \widehat{\Phi}_\ell(\widehat{v}_j)$. Thus the model prediction on the j-th block is $\widehat{\Gamma}_\ell[\widehat{v}_j]$. For $j = 0, \ldots, \kappa - 1$, define the block-wise risk by

$$L_j^{\mathrm{block}}[\widehat{\Gamma}_\ell] := \mathbb{E}_{u_0 \sim \Pi_\kappa,\, z \sim \mu}\left[\left|\widehat{\Gamma}_\ell[\widehat{v}_j](z) - \mathcal{G}_*[v_j](z)\right|^2\right].$$

Let $\mathcal{V}_\kappa(\widehat{\Gamma}_\ell) := \{v_j(u_0), \widehat{v}_j(u_0) : u_0 \in U_0^{[\kappa]},\ j = 0, \ldots, \kappa\} \subset U_{0,R}$, and define the rollout-relevant local error by

$$\varepsilon_{\mathrm{loc},\ell}^{[\kappa]} := \sup_{v \in \mathcal{V}_\kappa(\widehat{\Gamma}_\ell)} \left\|\widehat{\Gamma}_\ell[v] - \mathcal{G}_*[v]\right\|_{L^\infty(Q_T)}. \tag{10}$$

Theorem 6.1 (Block-wise long-time error propagation)**.** *Suppose that Assumption 3.1 holds. Then, for every* $j = 0, \ldots, \kappa - 1$*,*

$$L_j^{\mathrm{block}}[\widehat{\Gamma}_\ell] \leqslant \left(\sum_{r=0}^{j}\left(L_\psi \frac{C_S}{1-\delta}\right)^r\right)^2 \left(\varepsilon_{\mathrm{loc},\ell}^{[\kappa]}\right)^2.$$

This result shows that long-time prediction is obtained by deterministic propagation of the local error. Since no additional empirical risk minimization is performed after $\widehat{\Gamma}_\ell$ has been learned, the rollout

step does not introduce a new Rademacher complexity term. The block index j enters only through the stability factor in Theorem 6.1. As in Theorem 5.1, Theorem 6.1 also applies to any admissible equation class Ω_{eq} satisfying Assumption 3.1 (see also Remark 5.2). The proof iterates the local stability estimate (4) along the exact and approximate trajectories. See Appendix E for more details and the proof and also Appendix G for additional remarks.

7 A Concrete Example: Nonlinear Heat Equations on the Torus

We consider the nonlinear heat equation on $\mathbb{T}^d := \mathbb{R}^d/(2\pi\mathbb{Z})^d$:

$$\partial_t u - \Delta_{\mathbb{T}^d} u = F(u), \quad u(0) = u_0. \tag{11}$$

Let $R, M, L > 0$, and set $\mathcal{F}_{M,L} := \{F : [-M, M] \to \mathbb{R} : F(0) = 0,\ \mathrm{Lip}_{[-M,M]}(F) \leqslant L\}$ and $\Omega_{\mathrm{eq}} := \{(-\Delta_{\mathbb{T}^d}, F) : F \in \mathcal{F}_{M,L}\}$. For $F \in \mathcal{F}_{M,L}$ and $u_0 \in U_0$, we denote by $\mathcal{T}_{F,u_0}$ the corresponding Picard map. We assume that $R, M, L, T > 0$ and $\delta \in (0, 1)$ satisfy

$$R + TLM \leqslant M, \quad TL \leqslant \delta. \tag{12}$$

Then Assumption 3.1 holds with $C_S = 1$, using $C(\mathbb{T}^d)$ as the space of continuous functions on $\mathbb{T}^d$. Let $s_0 > d/2$ and $U_0 := \{u_0 \in H^{s_0}(\mathbb{T}^d) : \|u_0\|_{H^{s_0}} \leqslant R_0\} \subset U_{0,R}$. See Appendices B and F.1 for the details and the learning setting. We discuss only the full-information case here.

The linear heat semigroup is diagonalized by the Fourier basis: $e^{t\Delta_{\mathbb{T}^d}} f = \sum_{\xi\in\mathbb{Z}^d} e^{-t|\xi|^2} \widehat{f}(\xi) e_\xi$, where $e_\xi(x) := (2\pi)^{-d/2} e^{i\xi\cdot x}$. For the implementation, we use the tensor-product Fejér operator P_N and put

$$S_N(t)f := P_N e^{t\Delta_{\mathbb{T}^d}} f, \quad \mathcal{K}_N[g](t) := \int_0^t S_N(t-\tau) g(\tau)\, d\tau.$$

The operator P_N is L^∞-contractive and retains the modes $\xi \in \{-N, \dots, N\}^d$; see Appendix F.2.

Definition 7.1 (Picard-type Fourier neural operator)**.** Let $N, \ell, H \in \mathbb{N}$, and let $\rho : \mathbb{R} \to \mathbb{R}$ be a one-dimensional ReLU network of size at most H satisfying $\rho(0) = 0$ and $\mathrm{Lip}_{[-M,M]}(\rho) \leqslant L$. For $u_0 \in U_0$, define $u^{(0)}_{N,\rho}[u_0] := 0$ and

$$u^{(j+1)}_{N,\rho}[u_0] := S_N(\cdot)u_0 + \mathcal{K}_N\left[\rho\big(u^{(j)}_{N,\rho}[u_0]\big)\right], \quad j = 0, \dots, \ell - 1.$$

The ℓ-step Picard-type FNO $\Gamma^{\mathrm{FNO}}_{N,\ell,\rho} : U_0 \to U_M$ is defined by $\Gamma^{\mathrm{FNO}}_{N,\ell,\rho}(u_0) := u^{(\ell)}_{N,\rho}[u_0]$.

We denote the corresponding class by $\mathfrak{N}^{\mathrm{FNO}}_{N,\ell,H}(M, L)$. Only ρ is trainable and is shared across all Picard steps. Figure 1 illustrates the Picard-type FNO architecture. See Appendix F.2 for more details.

Theorem 7.2 (Implementation error for the Picard-type FNO)**.** *Assume* (12)*. Then, for any $\eta \in (0, 1)$ and any $F \in \mathcal{F}_{M,L}$, there exists a one-dimensional ReLU network ρ_F such that*

$$\|F - \rho_F\|_{L^\infty([-M,M])} \leqslant \eta, \quad \rho_F(0) = 0, \quad \mathrm{Lip}_{[-M,M]}(\rho_F) \leqslant L, \quad H(\rho_F) \leqslant C(1 + ML\eta^{-1}).$$

Moreover, there exist errors $a_N(\ell)$, independent of η, such that

$$\sup_{F\in\mathcal{F}_{M,L}} \sup_{u_0\in U_0} \left\| \Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}(u_0) - \mathcal{T}^{[\ell]}_{F,u_0}[0] \right\|_{L^\infty(Q_T)} \leqslant \frac{1-\delta^\ell}{1-\delta}\, T\eta + a_N(\ell).$$

In addition, for any $0 < \gamma < \min\{s_0 - d/2, 1\}$, there exists a constant $C_\gamma > 0$, independent of N, ℓ, η, such that

$$a_N(\ell) \leqslant C_\gamma N^{-\gamma} \quad \textit{for any } N, \ell \in \mathbb{N}.$$

Consequently, the implementation map $\mathrm{embed}_{N,\eta}(b_{(-\Delta_{\mathbb{T}^d},F),\ell}) := \Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}$*, defined using one fixed representative F for each predictor, satisfies*

$$\varepsilon^{\mathrm{FNO}}_{\mathrm{imp}}(N, \ell, \eta) \leqslant 4M \left\{ \frac{1-\delta^\ell}{1-\delta}\, T\eta + a_N(\ell) \right\}.$$

In particular, this error can be made arbitrarily small by choosing η sufficiently small and N sufficiently large, uniformly in ℓ.

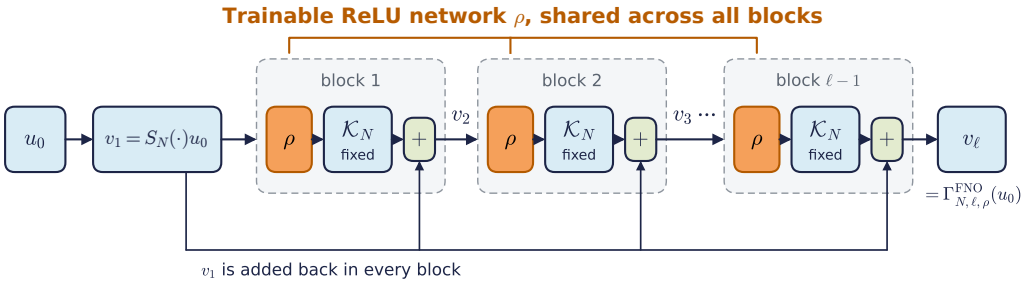


Figure 1: Picard-type FNO. The first iterate is $v_1 = S_N(\cdot)u_0$. Each subsequent block computes $v_{j+1} = v_1 + \mathcal{K}_N[\rho(v_j)]$, $j = 1, \dots, \ell - 1$. The same ReLU network ρ is shared across all blocks, while $\mathcal{K}_N$ is fixed. The term v_1 is added in every block. The output is $v_\ell = \Gamma^{\mathrm{FNO}}_{N,\ell,\rho}(u_0)$.

The proof uses piecewise affine approximation of F by a ReLU network, Picard stability, and uniform Fejér approximation on compact families. The explicit rate follows from Sobolev regularity and heat-semigroup smoothing; see Appendix F.3 and Proposition F.6.

Lemma 7.3 (Rademacher complexity of the Picard class)**.** *Assume* (12). *Then, for any $\ell \in \mathbb{N}$, the empirical Rademacher complexity of the Picard class satisfies*

$$\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} = \widehat{\mathfrak{R}}_{\mathcal{S}^{U_0}_{n,q}}\big(\mathcal{B}^{\Omega_{\mathrm{eq}}}_\ell\big) \leqslant CM\sqrt{\frac{LT(1-\delta^\ell)}{(1-\delta)n}},$$

where $C > 0$ is a universal constant.

The proof is given in Appendix F.4.

Theorem 7.4 (Generalization error bound for the Picard-type FNO)**.** *Assume* (12). *Let* $\widehat{b}_\ell \in \operatorname{argmin}_{b \in \mathcal{B}^{\Omega_{\mathrm{eq}}}_\ell} \widehat{L}_{n,q}[b]$ *and set* $\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta} := \mathrm{embed}_{N,\eta}(\widehat{b}_\ell)$. *Then there exists a universal constant $C > 0$ such that, for any $0 < \zeta < 1$, with probability at least $1 - \zeta$,*

$$L\Big[\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}\Big] \leqslant 4M\left\{\frac{1-\delta^\ell}{1-\delta}\,T\eta + a_N(\ell)\right\} + \frac{M^2\delta^{2\ell}}{(1-\delta)^2} + CM^2\sqrt{\frac{LT(1-\delta^\ell)}{(1-\delta)n}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}.$$

The statistical term is independent of the Fourier cutoff and remains bounded as the Picard depth increases. The proof and sufficient parameter choices for a prescribed population risk are given in Appendix F.5 and Corollary F.10. Long-time prediction by local rollout is treated in Appendix F.6. See also Appendix G for further discussion.

8 Conclusion

We developed a generalization theory for Picard-type operator learning for nonlinear parabolic PDEs. The main point is to use the Duhamel principle and Picard iteration as an abstract state-transition model and to separate the statistical complexity of this abstract Picard class from the approximation capability of a concrete neural-operator implementation. The resulting bounds decompose the error into implementation, Picard truncation, estimation, and finite-observation reconstruction terms. Under the contraction assumptions used throughout the paper, increasing the Picard depth improves the truncation error without increasing the entropy-based Rademacher term. We also showed how the same local view extends to rollout over multiple time blocks and how the theory can be instantiated for nonlinear heat equations on the torus with Picard-type FNOs.

Acknowledgments

This work was supported by JSPS KAKENHI 24K21316, 26K17014, JST BOOST JPMJBY24E2, JST CREST JPMJCR25I5, JST ERATO JPMJER2601, JST ARiSE JPMJAR2633.

References

B. Adcock, N. Dexter, and S. Moraga. Optimal deep learning of holomorphic operators between Banach spaces. In *Advances in Neural Information Processing Systems*, volume 37, pages 27725–27789, 2024.

S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

S. Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

V. Barbu. *Nonlinear Differential Equations of Monotone Types in Banach Spaces*. Springer Monographs in Mathematics. Springer, 2010.

M. Ben-Artzi, P. Souplet, and F. B. Weissler. The local theory for viscous Hamilton–Jacobi equations in Lebesgue spaces. *Journal de Mathématiques Pures et Appliquées*, 81(4):343–378, 2002.

K. Bhattacharya, B. Hosseini, N. B. Kovachki, and A. M. Stuart. Model reduction and neural networks for parametric PDEs. *SMAI Journal of Computational Mathematics*, 7:121–157, 2021.

B. Bonev, T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar. Spherical Fourier neural operators: Learning stable dynamics on the sphere. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2806–2823, 2023.

T. A. Bui, P. D'Ancona, X. T. Duong, and D. Müller. On the flows associated to selfadjoint operators on metric measure spaces. *Mathematische Annalen*, 375(3–4):1393–1426, 2019.

T. A. Bui, P. D'Ancona, and F. Nicola. Sharp L^p estimates for Schrödinger groups on spaces of homogeneous type. *Revista Matemática Iberoamericana*, 36(2):455–484, 2020.

M. Cannone. Harmonic Analysis Tools for Solving the Incompressible Navier–Stokes Equations. In S. Friedlander and D. Serre, editors, *Handbook of Mathematical Fluid Dynamics*, volume 3, pages 161–244. Elsevier, Amsterdam, 2004.

T. Cazenave. *Semilinear Schrödinger Equations*, volume 10 of *Courant Lecture Notes in Mathematics*. American Mathematical Society, Providence, RI, 2003.

T. Cazenave and A. Haraux. *An Introduction to Semilinear Evolution Equations*, volume 13. Oxford University Press, Oxford, 1998.

G. Chen, X. Liu, Q. Meng, L. Chen, C. Liu, and Y. Li. Learning neural operators on Riemannian manifolds. *National Science Open*, 3(6):20240001, 2024.

K. Chen, C. Wang, and H. Yang. Deep operator learning lessens the curse of dimensionality for PDEs. *Transactions on Machine Learning Research*, 2023.

T. Chen and H. Chen. Approximations of continuous functionals by neural networks with application to dynamic systems. *IEEE Transactions on Neural Networks*, 4(6):910–918, 1993.

T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6:911–917, 1995.

M. G. Crandall and T. M. Liggett. Generation of semi-groups of nonlinear transformations on general Banach spaces. *American Journal of Mathematics*, 93(2):265–298, 1971.

E. B. Davies. *Heat Kernels and Spectral Theory*, volume 92 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

E. B. Davies. Heat kernel bounds for higher order elliptic operators. *Journées Équations aux Dérivées Partielles*, pages 1–11, 1995.

B. Deng, Y. Shin, L. Lu, Z. Zhang, and G. E. Karniadakis. Approximation rates of DeepONets for learning operators arising from advection–diffusion equations. *Neural Networks*, 153:411–426, 2022.

P. Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, volume 35 of *Springer Series in Computational Mathematics*. Springer, Berlin and Heidelberg, 2011.

M. Feischl, C. Schwab, and F. Zehetgruber. Neural general operator networks via Banach fixed point iterations. Technical Report SAM Report 2025-13, Seminar for Applied Mathematics, ETH Zürich, Zürich, Switzerland, 2025.

T. Furuya and A. Kratsios. Polynomial scaling is possible for neural operator approximations of structured families of BSDEs, 2024. arXiv preprint arXiv:2410.14788.

T. Furuya, K. Taniguchi, and S. Okuda. Quantitative approximation for neural operators in nonlinear parabolic equations. In *Proceedings of the International Conference on Learning Representations*, 2025.

Y. Giga and T. Miyakawa. Solutions in L^r of the Navier–Stokes initial value problem. *Archive for Rational Mechanics and Analysis*, 89(3):267–281, 1985.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2015. doi: 10.1017/CBO9781107337862.

P. Gopalani, S. Karmakar, D. Kumar, and A. Mukherjee. Towards size-independent generalization bounds for deep operator nets, 2022.

L. Grafakos. *Classical Fourier Analysis*, volume 249 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 3 edition, 2014.

G. Gupta, X. Xiao, and P. Bogdan. Multiwavelet-based operator learning for differential equations. In *Advances in Neural Information Processing Systems*, volume 34, pages 24048–24062, 2021.

I. Henderson. Sobolev regularity of Gaussian random fields. *Journal of Functional Analysis*, 286(3): 110241, 2024. doi: 10.1016/j.jfa.2023.110241.

D. Henry. *Geometric Theory of Semilinear Parabolic Equations*, volume 840 of *Lecture Notes in Mathematics*. Springer, Berlin, 1981.

L. Herrmann, C. Schwab, and J. Zech. Neural and spectral operator surrogates: Unified construction and expression rate bounds. *Advances in Computational Mathematics*, 50:72, 2024.

M. Ikeda, K. Taniguchi, and Y. Wakasugi. Global existence and asymptotic behavior for semilinear damped wave equations on measure spaces. *Evolution Equations and Control Theory*, 13(4): 1101–1125, 2024.

T. Iwabuchi, T. Matsuyama, and K. Taniguchi. Boundedness of spectral multipliers for Schrödinger operators on open sets. *Revista Matemática Iberoamericana*, 34(3):1277–1322, 2018.

B. Kaltenbacher and K. V. Huynh. Iterative regularization for constrained minimization formulations of nonlinear inverse problems. *Computational Optimization and Applications*, 81:569–611, 2022.

Y. Katznelson. *An Introduction to Harmonic Analysis*. Cambridge University Press, 3 edition, 2004.

T. Kim and M. Kang. Bounding the Rademacher complexity of Fourier neural operators. *Machine Learning*, 113:2467–2498, 2024.

N. Kovachki, S. Lanthaler, and S. Mishra. On universal approximation and error bounds for Fourier neural operators. *Journal of Machine Learning Research*, 22:1–76, 2021.

N. B. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. M. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24:1–97, 2023.

N. B. Kovachki, S. Lanthaler, and H. Mhaskar. Data complexity estimates for operator learning, 2024a. arXiv preprint arXiv:2405.15992.

N. B. Kovachki, S. Lanthaler, and A. M. Stuart. Operator learning: Algorithms and analysis. *Handbook of Numerical Analysis*, 25:419–467, 2024b.

S. Lanthaler. Operator learning with PCA-Net: Upper and lower complexity bounds. *Journal of Machine Learning Research*, 24:1–67, 2023.

S. Lanthaler and A. M. Stuart. The parametric complexity of operator learning. *IMA Journal of Numerical Analysis*, 46(2):647–712, 2026.

S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for DeepONets: A deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6:tnac001, 2022.

S. Lanthaler, Z. Li, and A. M. Stuart. Nonlocality and nonlinearity implies universality in operator learning. *Constructive Approximation*, 62(2):261–303, 2025.

J. A. Lara Benitez, T. Furuya, F. Faucher, A. Kratsios, X. Tricoche, and M. V. de Hoop. Out-of-distributional risk bounds for neural operators with applications to the Helmholtz equation. *Journal of Computational Physics*, 513:113168, 2024.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. M. Stuart, and A. Anandkumar. Neural operator: Graph kernel network for partial differential equations, 2020. arXiv preprint arXiv:2003.03485.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. M. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *Proceedings of the International Conference on Learning Representations*, 2021.

P. Lippe, B. S. Veeling, P. Perdikaris, R. E. Turner, and J. Brandstetter. PDE-refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3): 218–229, 2021.

C. Marcati and C. Schwab. Exponential convergence of deep operator networks for elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 61(3):1513–1545, 2023.

C. Marcati and C. Schwab. Expression rates of neural operators for linear elliptic PDEs in polytopes, 2024. arXiv preprint arXiv:2409.17552.

T. Marwah, A. Pokle, J. Z. Kolter, Z. C. Lipton, J. Lu, and A. Risteski. Deep equilibrium based neural operators for steady-state PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pages 15716–15737, 2023.

A. Maurer. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory*, volume 9925 of *Lecture Notes in Computer Science*, pages 3–17. Springer, 2016.

K. Michałowska, S. Goswami, G. E. Karniadakis, and S. Riemer-Sørensen. Neural operator learning for long-time integration in dynamical systems with recurrent neural networks, 2023. arXiv preprint arXiv:2303.02243.

T. T. N. Nguyen. Bi-level iterative regularization for inverse problems in nonlinear PDEs. *Inverse Problems*, 40(4):045020, 2024.

E. M. Ouhabaz. *Analysis of Heat Equations on Domains*, volume 31 of *London Mathematical Society Monographs*. Princeton University Press, Princeton, NJ, 2005.

C. V. Pao. *Nonlinear Parabolic and Elliptic Equations*. Springer, New York, 1992.

A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*, volume 44 of *Applied Mathematical Sciences*. Springer, New York, 1983.

C. Schwab, A. Stein, and J. Zech. Deep operator network approximation rates for Lipschitz operators. *Analysis and Applications*, 24(1):199–239, 2026.

S. Sonoda, Y. Hashimoto, I. Ishikawa, and M. Ikeda. Why and when deep is better than shallow: An implementation-agnostic state-transition view of depth supremacy, 2025. arXiv preprint arXiv:2505.15064.

T. Tao. *Nonlinear Dispersive Equations: Local and Global Analysis*, volume 106 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 2006.

T. Tripura and S. Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

E. Zeidler. *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems*. Springer, New York, 1986.

Limitations

The analysis is theoretical and relies on structural assumptions on the target equation class, most notably local well-posedness, uniform boundedness, and a contractive Picard map on the time interval under consideration. The implementation error is kept abstract in the main results, and it must be estimated separately for each concrete architecture and PDE class; we provide one such illustration for the nonlinear heat equation on the torus. The long-time result also depends on stability of the local solution map and on the rollout staying inside the admissible state set. These assumptions are standard for this type of analysis, but they limit the direct scope of the bounds for strongly unstable dynamics or for settings where the learned model is used far outside the training distribution.

Broader Impacts

This work is foundational and does not introduce a deployed system, dataset, or high-risk model. Its main potential benefit is to clarify when solver-aligned operator learning can provide reliable surrogate models for PDE-driven scientific and engineering problems, which may reduce the cost of repeated numerical simulation. A possible negative impact is overreliance on learned surrogates in safety-critical physical modeling when the assumptions behind the generalization bound are not checked. In practical deployments, the bounds should therefore be complemented by problem-specific validation, uncertainty assessment, and monitoring of whether inputs remain within the regime covered by the model assumptions.

A Related Work

Operator learning architectures. Operator learning models include DeepONet [Lu et al., 2021, Lanthaler et al., 2022], PCA-Net [Bhattacharya et al., 2021, Lanthaler, 2023], graph neural operators [Li et al., 2020], Fourier neural operators [Li et al., 2021, Kovachki et al., 2021], wavelet and multiwavelet neural operators [Gupta et al., 2021, Tripura and Chakraborty, 2023], and recent variants such as geometric or equivariant neural operators [Bonev et al., 2023, Chen et al., 2024]. Universal approximation results are available for several of these architectures [Chen and Chen, 1993, 1995, Kovachki et al., 2021, Lanthaler et al., 2022, Lanthaler, 2023, Lanthaler et al., 2025]. These results establish expressivity for broad operator classes, but they do not by themselves explain when an operator can be learned efficiently from finite data or approximated without an exponential dependence on the ambient function-space dimension.

Approximation theory and structural PDE classes. For general continuous or Lipschitz operators, operator learning can suffer from the curse of parametric complexity [Lanthaler and Stuart, 2026]. A common way to avoid this obstruction is to exploit additional structure of the target operator. Positive approximation results have been obtained for solution operators of Darcy flow, Navier–Stokes equations, elliptic and parabolic problems, advection–diffusion equations, Hamilton–Jacobi equations, and holomorphic operator classes [Kovachki et al., 2021, Chen et al., 2023, Lanthaler, 2023, Deng et al., 2022, Marcati and Schwab, 2023, Schwab et al., 2026, Furuya and Kratsios, 2024, Marcati and Schwab, 2024, Herrmann et al., 2024, Adcock et al., 2024, Lanthaler and Stuart, 2026]. These works show that efficient operator approximation is possible when the target class has exploitable regularity, spectral structure, low-dimensional compression, holomorphy, characteristics, or other PDE-specific structure.

Generalization and capacity in operator learning. Finite-sample analysis for operator learning has been developed for DeepONets, FNO-type models, and more general operator classes [Lanthaler et al., 2022, Gopalani et al., 2022, Kim and Kang, 2024, Lara Benitez et al., 2024, Kovachki et al., 2024a]. These works typically analyze either the capacity of a concrete architecture or the sample complexity of a specified operator approximation class. Our focus is complementary. We do not estimate the Rademacher complexity of all parameters in a neural operator. Instead, following the implementation-agnostic perspective of Sonoda et al. [2025], we analyze the statistical complexity of the abstract Picard transition class that the implementation is intended to realize. The concrete architecture enters through an implementation error.

Picard-aligned and fixed-point neural operators. The closest approximation-theoretic predecessor is Furuya et al. [2025]. They rewrite nonlinear parabolic PDEs in Duhamel form and align each forward layer of a neural operator with one Picard step. Their result shows that the solution operator can be approximated without exponential growth in model complexity; the neural network component is used mainly to approximate the scalar nonlinearity. This gives the approximation side of the present work. We add the statistical side: once the Picard steps are viewed as a contractive state-transition system, the Rademacher term can be controlled through the entropy of the transition family and need not grow with Picard depth.

This perspective is also related to fixed-point models in machine learning, including deep equilibrium models [Bai et al., 2019], FNO-DEQ models for steady-state PDEs [Marwah et al., 2023], and neural general operator networks based on constructive fixed-point principles [Feischl et al., 2025]. Those works incorporate fixed-point structure into architectures or approximation schemes. Our contribution is a finite-sample generalization analysis for the abstract solution class generated by Duhamel–Picard iteration.

Depth, state transitions, and rollout. Sonoda et al. [2025] studies when depth improves generalization in an implementation-agnostic state-transition model. Section 6.2 of that work identifies Picard-aligned neural operators as a favorable regime: the target solution operator has an iterative structure, Picard approximation improves rapidly with depth, and contractive transitions keep word-ball entropy controlled. The present paper instantiates this observation for nonlinear parabolic PDEs, adds finite-observation reconstruction errors, and derives operator-learning generalization bounds. For long-time prediction, related experimental work studies error accumulation and rollout stabilization in neural PDE solvers [Michałowska et al., 2023, Lippe et al., 2023]. Our rollout analysis is a deterministic propagation result for a learned local solution operator; it does not add a new statistical complexity term after the local model has been learned.

B Sufficient Conditions and Examples of Admissible Equation Class Ω_{eq}

In this appendix, we give a sufficient condition for Assumption 3.1, together with examples of linear operators A, nonlinearities F, and admissible equation classes $\Omega_{\mathrm{eq}} := \mathcal{A} \times \mathcal{F}$. We also explain how a common local time can be chosen and discuss the scope of the present setting of solution space.

To begin with, let us recall Assumption 3.1 of admissible equation class Ω_{eq}: there exist $R, M, T, C_S > 0$ and $\delta \in (0,1)$ such that the following conditions (a) and (b) hold.

(a) The map $\mathcal{T}_{\omega,u_0} : U_M \to U_M$ is δ-contractive, i.e.,

$$\|\mathcal{T}_{\omega,u_0}[u] - \mathcal{T}_{\omega,u_0}[v]\|_{L^\infty(Q_T)} \leqslant \delta \|u - v\|_{L^\infty(Q_T)} \tag{13}$$

for any $u, v \in U_M$.

(b) The map $\mathcal{T}_{\omega,u_0}$ is Lipschitz continuous with respect to u_0, i.e.,

$$\|\mathcal{T}_{\omega,u_0}[u] - \mathcal{T}_{\omega,v_0}[u]\|_{L^\infty(Q_T)} \leqslant C_S \|u_0 - v_0\|_{L^\infty(D)} \tag{14}$$

for any $u_0, v_0 \in U_{0,R}$ and $u \in U_M$.

The conditions (a) and (b) in Assumption 3.1 are the standard conditions used in Banach's fixed-point theorem to obtain local well-posedness [Banach, 1922, Henry, 1981, Pazy, 1983, Zeidler, 1986]. Indeed, Assumption 3.1 yields Proposition 3.2, i.e., the result on local well-posedness for (P) in $C([0,T]; C(D))$.

B.1 A sufficient condition for Assumption 3.1

We state a typical and useful sufficient condition on Ω_{eq} that guarantees Assumption 3.1.

Proposition B.1. *Let $R, M, T > 0$ and $\delta \in (0,1)$. Set $C_b(D) := C(D) \cap L^\infty(D)$, equipped with the supremum norm. Let $\mathcal{A}$ be a class of linear operators with associated C_0-semigroups $S_A(t)$ on $L^2(D)$ such that, for any $A \in \mathcal{A}$, the restriction of $S_A(t)$ to $C_b(D)$ is a C_0-semigroup on $C_b(D)$, and*

$$C_{\mathcal{A}}(T) := \sup_{A \in \mathcal{A}} \sup_{0 < t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} < \infty. \tag{15}$$

Let $\mathcal{F}$ be a class of functions $F : \mathbb{R} \to \mathbb{R}$ satisfying $F(0) = 0$ and

$$L_{\mathcal{F}}(M) := \sup_{F \in \mathcal{F}} \mathrm{Lip}_{[-M,M]}(F) < \infty. \tag{16}$$

If

$$C_{\mathcal{A}}(T)R + TC_{\mathcal{A}}(T)L_{\mathcal{F}}(M)M \leqslant M, \quad TC_{\mathcal{A}}(T)L_{\mathcal{F}}(M) \leqslant \delta, \tag{17}$$

then $\Omega_{\mathrm{eq}} := \mathcal{A} \times \mathcal{F}$ satisfies Assumption 3.1 with $C_S = C_{\mathcal{A}}(T)$.

Proof. The inequality (14) immediately follows from (15). In fact, by the definition of the Picard map, we have

$$\mathcal{T}_{\omega,u_0}[u](t) - \mathcal{T}_{\omega,v_0}[u](t) = S_A(t)(u_0 - v_0)$$

for $t \in (0, T)$, and hence,

$$\begin{aligned}\|\mathcal{T}_{\omega,u_0}[u] - \mathcal{T}_{\omega,v_0}[u]\|_{L^\infty(Q_T)} &\leqslant \sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} \|u_0 - v_0\|_{L^\infty(D)} \\ &\leqslant C_{\mathcal{A}}(T)\|u_0 - v_0\|_{L^\infty(D)}\end{aligned}$$

This implies that the condition (b) holds.

Next, we show that the condition (a) holds. Now, an operator $A \in \mathcal{A}$ generates a C_0-semigroup on $L^2(D)$ and it satisfies the L^∞-estimate (15). Moreover, the nonlinear term $F : \mathbb{R} \to \mathbb{R}$ in $\mathcal{F}$ satisfies $F(0) = 0$ and local Lipschitz condition in U_M for $M > 0$, i.e., there exists a constant $\mathrm{Lip}_{[-M,M]}(F) > 0$ such that

$$|F(z_1) - F(z_2)| \leqslant \mathrm{Lip}_{[-M,M]}(F)|z_1 - z_2| \tag{18}$$

for any $z_1, z_2 \in [-M, M]$. We take R, M, T and $\delta \in (0, 1)$ as

$$\sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} R + T \sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} \mathrm{Lip}_{[-M,M]}(F)M \leqslant M \tag{19}$$

and

$$T \sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} \mathrm{Lip}_{[-M,M]}(F) \leqslant \delta. \tag{20}$$

For each $u_0 \in U_{0,R}$ and for any $u \in U_M$, it follows from $F(0) = 0$, (18) and (19) that

$$\begin{aligned}\|\mathcal{T}_{\omega,u_0}[u]\|_{L^\infty(Q_T)} &\leqslant \|S_A(t)u_0\|_{L^\infty(Q_T)} + \left\| \int_0^t S_A(t-\tau)F(u(\tau))\,\tau \right\|_{L^\infty(Q_T)} \\ &\leqslant \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} \|u_0\|_{L^\infty(D)} \\ &\quad + \sup_{0 < t \leqslant T} \int_0^t \|S_A(t-\tau)\|_{L^\infty(D) \to L^\infty(D)}\, d\tau \times \mathrm{Lip}_{[-M,M]}(F)\|u\|_{L^\infty(Q_T)} \\ &\quad \sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} R \\ &\quad + T \sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} \mathrm{Lip}_{[-M,M]}(F)M \leqslant M.\end{aligned}$$

Moreover, since $S_A(t)$ is the C_0-semigroup, $\mathcal{T}_{\omega,u_0}[u]$ is continuous in Q_T by the standard argument. Hence, it is shown that $\mathcal{T}_{\omega,u_0}$ is a map from U_M into itself. Next, for each $u_0 \in U_{0,R}$ and for any $u, v \in U_M$, we write

$$\mathcal{T}_{\omega,u_0}[u](t) - \mathcal{T}_{\omega,u_0}[v](t) = \int_0^t S_A(t-\tau)\big(F(u(\tau)) - F(v(\tau))\big)\, d\tau.$$

Since $u, v \in U_M$, we have $u(\tau, x), v(\tau, x) \in [-M, M]$ for a.e. $(\tau, x) \in Q_T$. Hence, it follows from (18) and (20) that

$$\begin{aligned}&\|\mathcal{T}_{\omega,u_0}[u] - \mathcal{T}_{\omega,u_0}[v]\|_{L^\infty(Q_T)} \\ &\leqslant \sup_{0 < t \leqslant T} \int_0^t \|S_A(t-\tau)\|_{L^\infty(D) \to L^\infty(D)}\, d\tau \times \mathrm{Lip}_{[-M,M]}(F)\|u - v\|_{L^\infty(Q_T)} \\ &\leqslant T \sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} \mathrm{Lip}_{[-M,M]}(F)\|u - v\|_{L^\infty(Q_T)} \\ &\leqslant \delta\|u - v\|_{L^\infty(Q_T)}.\end{aligned}$$

This proves (13). Since $S_A(t)$ is strongly continuous on $C_b(D)$ and F is Lipschitz continuous on $[-M, M]$, the Duhamel formula gives $\mathcal{T}_{\omega,u_0}[u] \in C([0, T]; C_b(D))$. Together with the above norm estimate, this shows that $\mathcal{T}_{\omega,u_0}[u] \in U_M$. Summarizing the above, we conclude Proposition B.1. □

Remark B.2 (Choice of a common local time). Fix $R, M > 0$ and $\delta \in (0,1)$. Suppose that

$$C_{\mathcal{A}}(T_0) \leqslant C_S, \quad L_{\mathcal{F}}(M) \leqslant L, \quad C_S R < M$$

for some $T_0, C_S, L > 0$. Then (17) holds whenever

$$0 < T \leqslant \min\left\{T_0, \frac{M - C_S R}{C_S L M}, \frac{\delta}{C_S L}\right\}. \tag{21}$$

Thus, these common bounds give a common local time for all equations in the class. If $L_{\mathcal{F}}(M) = 0$, any $0 < T \leqslant T_0$ is allowed. If the semigroups are L^∞-contractive, we can take $C_S = 1$, so the condition $C_S R < M$ becomes $R < M$.

Remark B.3 (Continuity and boundary compatibility). The L^∞-bound (15) and strong continuity on $L^2(D)$ alone do not ensure the continuity required by U_M. This is why Proposition B.1 also assumes strong continuity on $C_b(D)$. One may replace $C_b(D)$ by a closed subspace of bounded continuous functions, equipped with the supremum norm, on which the semigroups are strongly continuous and which is preserved by $u \mapsto F(u)$. For the torus, one uses $C(\mathbb{T}^d)$. For homogeneous Dirichlet problems on a sufficiently regular bounded domain, one uses

$$C_0(D) := \{f \in C(\overline{D}) : f|_{\partial D} = 0\}.$$

Since $F(0) = 0$, the nonlinearity preserves $C_0(D)$. The initial-data and solution spaces must then be defined using this space, and the sampling distribution must be supported on initial data satisfying the boundary condition.

B.2 Examples of linear operators A

In this appendix, we give examples of linear operators A satisfying

$$\sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} < \infty. \tag{22}$$

As shown in the proof of Proposition B.1, (22) is equivalent to the condition (b) with $C_S = \sup_{0 \leqslant t \leqslant T} \|S_A(t)\|_{L^\infty(D) \to L^\infty(D)}$. For simplicity, and to keep the setting consistent with this paper, we state the examples only for A on bounded domains D. However, it is known that (22) also holds in many cases for A on the whole Euclidean spaces $\mathbb{R}^d$, unbounded domains, Riemannian manifolds, groups, fractals or general metric measure spaces (see, e.g., [Bui et al., 2019, 2020, Davies, 1989, Furuya et al., 2025, Ikeda et al., 2024, Ouhabaz, 2005] and references therein).

We list several examples below.

1. **(Laplacian on the torus)** Let $A = -\Delta_{\mathbb{T}^d}$ be the Laplacian on $\mathbb{T}^d := \mathbb{R}^d/(2\pi\mathbb{Z})^d$, that is, the Laplacian on $D = (0, 2\pi)^d$ with periodic boundary condition. Then $S_A(t) = e^{t\Delta_{\mathbb{T}^d}}$ is diagonalized by the Fourier basis

   $$e_\xi(x) := (2\pi)^{-d/2} e^{i\xi \cdot x}, \quad \xi \in \mathbb{Z}^d.$$

   More precisely, we have

   $$S_A(t) f = \sum_{\xi \in \mathbb{Z}^d} e^{-t|\xi|^2} \widehat{f}(\xi) e_\xi,$$

   where $\widehat{f}$ is the Fourier coefficient of f:

   $$\widehat{f}(\xi) := \int_{\mathbb{T}^d} f(x) \overline{e_\xi(x)}\, dx.$$

   Moreover, the heat semigroup is L^∞-contractive:

   $$\sup_{t > 0} \|S_A(t)\|_{L^\infty(\mathbb{T}^d) \to L^\infty(\mathbb{T}^d)} \leqslant 1. \tag{23}$$

   Therefore, $A = -\Delta_{\mathbb{T}^d}$ satisfies (22) with $T = \infty$, and we can take $C_S = 1$ in (b).

2. **(Dirichlet Laplacian)** Let $D \subset \mathbb{R}^d$ be a bounded domain, and let $A = -\Delta_{\mathrm{Dir}}$ be the Dirichlet Laplacian on D. Then $S_A(t) = e^{t\Delta_{\mathrm{Dir}}}$ is L^∞-contractive and has an exponential time decay:

   $$\|S_A(t)\|_{L^\infty(D) \to L^\infty(D)} \leqslant \min\{1, Ce^{-\lambda_1 t}\}, \quad t > 0 \tag{24}$$

   for some constant $C \geqslant 1$, where λ_1 is the first eigenvalue of $-\Delta_{\mathrm{Dir}}$. Thus, $A = -\Delta_{\mathrm{Dir}}$ satisfies (22) with $T = \infty$, and we can take $C_S = 1$ in (b). See, for example, [Davies, 1989], [Furuya et al., 2025, Appendix A.1], [Ikeda et al., 2024, Subsection 5.1].

3. **(Neumann or Robin Laplacian)** Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let $A = -\Delta_\gamma$ be the Robin Laplacian on D associated with the boundary condition

$$\partial_\nu u + \gamma u = 0 \quad \text{on } \partial D,$$

where $\gamma \in L^\infty(\partial D)$ and $\gamma \geqslant 0$. The case $\gamma = 0$ corresponds to the Neumann Laplacian $A = -\Delta_{\mathrm{Neu}}$. Then $S_A(t) = e^{-tA}$ satisfies L^∞-contractive and has an exponential time decay (24), where one has $\lambda_1 = 0$ in the pure Neumann case $\gamma = 0$. Therefore, $A = -\Delta_\gamma$ satisfies (22) with $T = \infty$, and we can take $C_S = 1$ in (b). See, for example, [Ouhabaz, 2005], [Furuya et al., 2025, Appendix A.2].

4. **(Schrödinger operator with a singular potential)** Let $D \subset \mathbb{R}^d$ be a bounded domain, and let $A = -\Delta_{\mathrm{Dir}} + V$ be the Schrödinger operator with the zero Dirichlet boundary condition, where $V = V_+ - V_-$ is a real-valued measurable potential satisfying

$$V_\pm \geqslant 0, \quad V_+ \in L^1_{\mathrm{loc}}(D), \quad V_- \in K_d(D),$$

where $K_d(D)$ denotes the Kato class on D. Then the semigroup $S_A(t) = e^{-tA}$ satisfies the following L^∞-estimate: For each $T > 0$, there exists a constant $C_A(T) > 0$ such that

$$\sup_{0<t\leqslant T} \|S_A(t)\|_{L^\infty(D)\to L^\infty(D)} \leqslant C_A(T). \tag{25}$$

Hence, $A = -\Delta_{\mathrm{Dir}} + V$ satisfies (22), and we can take $C_S = C_A(T)$ in (b). If $V \geqslant 0$ (i.e. $V_- = 0$), then (23) holds. Therefore, $A = -\Delta_{\mathrm{Dir}} + V$ satisfies (22) with $T = \infty$, and we can take $C_S = 1$ in (b). See, for example, [Iwabuchi et al., 2018], [Furuya et al., 2025, Appendix A.3], [Ikeda et al., 2024, Subsection 5.3].

5. **(Schrödinger operator with a Dirac delta potential)** Let $D = (a, b) \subset \mathbb{R}$, $x_0 \in D$, and $\alpha > 0$. We consider the one-dimensional Schrödinger operator

$$A = -\frac{d^2}{dx^2} + \alpha\delta_{x_0}$$

with the zero Dirichlet boundary condition. More precisely, A is realized as the Laplacian on $D\backslash\{x_0\}$ with the interface condition $u'(x_0+) - u'(x_0-) = \alpha u(x_0)$. Then $S_A(t) = e^{-tA}$ satisfies the L^∞-contractive estimate (23). Hence, $A = -d^2/dx^2 + \alpha\delta_{x_0}$ satisfies (22) with $T = \infty$, and we can take $C_S = 1$ in (b). See, for example, [Ikeda et al., 2024, Subsection 5.4].

6. **(Uniformly elliptic operator)** Let $D \subset \mathbb{R}^d$ be a bounded domain, and let A be a uniformly elliptic operator of divergence form

$$Au = -\sum_{i,j=1}^d \partial_i\big(a_{ij}(x)\partial_j u\big) + c(x)u$$

with a suitable boundary condition, such as the Dirichlet, Neumann, or Robin boundary condition. We assume that $a_{ij} \in L^\infty(D)$, $c \in L^\infty(D)$, $c \geqslant 0$, and that there exists $C > 0$ such that

$$\sum_{i,j=1}^d a_{ij}(x)\xi_i\xi_j \geqslant C|\xi|^2$$

for all $\xi \in \mathbb{R}^d$ and a.e. $x \in D$. Then $S_A(t) = e^{-tA}$ satisfies the L^∞-contractive estimate (23) under the standard maximum principle assumptions. Hence, A satisfies (22) with $T = \infty$, and we can take $C_S = 1$ in (b). See, for example, [Ouhabaz, 2005], [Furuya et al., 2025, Appendix A.4], [Ikeda et al., 2024, Subsection 5.5] for more general cases.

7. **(Fractional order or higher order case)** We consider the fractional powers $A^{\alpha/2}$, $\alpha > 0$, of the above non-negative operators A. Then we can see that the semigroup $S_{A^{\alpha/2}}(t) = e^{-tA^{\alpha/2}}$ satisfies (23) or (25). Hence, $A^{\alpha/2}$ satisfies (22). See, for example, [Davies, 1995, Ouhabaz, 2005], [Ikeda et al., 2024, Subsection 5.7] for more details.

B.3 Examples of nonlinearities

The nonlinear part of Proposition B.1 requires $F(0) = 0$ and a uniform Lipschitz bound

$$L_{\mathcal{F}}(M) := \sup_{F\in\mathcal{F}} \mathrm{Lip}_{[-M,M]}(F) < \infty.$$

We list several examples below.

1. **(Power nonlinearities)** Let
$$F(u) = \lambda |u|^{p-1} u \quad (\lambda \in \mathbb{R},\ p \geqslant 1).$$
Then $F(0) = 0$ and
$$\mathrm{Lip}_{[-M,M]}(F) \leqslant |\lambda| p M^{p-1}.$$
Thus, for fixed p, a bounded set of coefficients λ gives a uniform Lipschitz bound. The exponent p may also vary in a bounded subset of $[1, \infty)$.

2. **(Polynomial nonlinearities)** Let
$$F(u) = \sum_{j=1}^{p} a_j u^j \quad (p \in \mathbb{N},\ a_j \in \mathbb{R}).$$
Then $F(0) = 0$ and
$$\mathrm{Lip}_{[-M,M]}(F) \leqslant \sum_{j=1}^{p} j |a_j| M^{j-1}.$$
Hence, uniformly bounded degrees and coefficients give a uniform Lipschitz bound.

3. **(Allen–Cahn and Fisher–KPP nonlinearities)** The nonlinearities
$$F(u) = u - u^3, \quad F(u) = u - u^2$$
satisfy $F(0) = 0$, with Lipschitz bounds $1 + 3M^2$ and $1 + 2M$, respectively. These are special cases of the polynomial nonlinearities above.

4. **(Other smooth nonlinearities)** The functions
$$F(u) = \sin u, \quad F(u) = \tanh u, \quad F(u) = e^u - 1$$
satisfy $F(0) = 0$, with Lipschitz bounds 1, 1, and e^M, respectively. More generally, any class $\mathcal{F} \subset C^1(\mathbb{R})$ satisfying $F(0) = 0$ for any $F \in \mathcal{F}$ and
$$\sup_{F \in \mathcal{F}} \sup_{|u| \leqslant M} |F'(u)| < \infty$$
has a uniform Lipschitz bound on $[-M, M]$.

Remark B.4 (Time- and space-dependent nonlinearities). The results in Sections 3–5 extend to nonlinearities $F = F(t, x, u)$ satisfying $F(t, x, 0) = 0$ and
$$\sup_{F \in \mathcal{F}} \sup_{(t,x) \in [0,T_0] \times D} \mathrm{Lip}_{[-M,M]}\big(F(t, x, \cdot)\big) < \infty$$
for some $T_0 > 0$, provided that the required continuity and boundary conditions are preserved and a common local time is chosen as in Proposition B.1. The equation class and the Picard map must be modified accordingly. The entropy-based bounds require the corresponding entropy assumptions on the modified class. For time-dependent equations, the rollout formulation in Section 6 must also account for the dependence of the local solution map on the starting time.

Remark B.5 (Scope of the criterion). The criterion applies to reaction terms $F = F(u)$ that are Lipschitz continuous on $[-M, M]$. It does not directly cover
$$F(u) = |u|^p \quad (0 < p < 1), \qquad F(u) = u \log |u|,$$
where the latter is defined to be zero at $u = 0$, since these functions are not locally Lipschitz at zero. Derivative-dependent terms, such as $|\nabla u|^p$ with $p > 1$, also lie outside the present setting. Such equations require separate analysis. To apply the present framework, one must choose suitable solution spaces and establish the required contraction and stability estimates.

B.4 Examples of admissible equation classes

We combine the preceding examples to construct admissible equation classes. In each case, we use the continuous function space specified below to define the initial-data and solution spaces. Choose $0 < R < M$, $L > 0$, $\delta \in (0, 1)$, and $T > 0$ such that
$$R + TLM \leqslant M, \quad TL \leqslant \delta.$$

1. **(Unknown nonlinearity)** Fix $A = -\Delta_{\mathbb{T}^d}$ and set

$$\mathcal{F}_{M,L} := \left\{ F : [-M, M] \to \mathbb{R} : F(0) = 0,\ \mathrm{Lip}_{[-M,M]}(F) \leqslant L \right\},$$

$$\Omega_{\mathrm{eq}} := \{(-\Delta_{\mathbb{T}^d}, F) : F \in \mathcal{F}_{M,L}\}.$$

Each F may be extended to $\mathbb{R}$ without increasing its Lipschitz constant. We use $C(\mathbb{T}^d)$ as the continuous function space. The heat semigroup is strongly continuous on this space and is L^∞-contractive. Thus, the assumptions of Proposition B.1 hold with $C_{\mathcal{A}}(T) = 1$ and $L_{\mathcal{F}}(M) \leqslant L$. Hence, Assumption 3.1 holds with $C_S = 1$.

Here, the unknown equation specification is the nonlinearity. This class is studied in Section 7 and Appendix F.1.

2. **(Unknown boundary condition)** Let D be a bounded smooth domain. Fix a locally Lipschitz function $F : \mathbb{R} \to \mathbb{R}$ with $F(0) = 0$ and $\mathrm{Lip}_{[-M,M]}(F) \leqslant L$. For a fixed $\gamma_0 > 0$, set

$$\mathcal{R}_{\gamma_0} := \{\gamma \in C^1(\partial D) : 0 \leqslant \gamma \leqslant \gamma_0\}, \quad \Omega_{\mathrm{eq}} := \{(-\Delta_\gamma, F) : \gamma \in \mathcal{R}_{\gamma_0}\},$$

where $-\Delta_\gamma$ is the Robin Laplacian with boundary condition

$$\partial_\nu u + \gamma u = 0 \quad \text{on } \partial D.$$

We use $C(\overline{D})$ as the continuous function space. The Robin heat semigroups are strongly continuous on this space and are L^∞-contractive. Thus, the assumptions of Proposition B.1 hold with $C_{\mathcal{A}}(T) = 1$ and $\mathcal{F} = \{F\}$. Hence, Assumption 3.1 holds with $C_S = 1$.

Here, the unknown equation specification is the boundary coefficient γ.

3. **(Unknown coefficients)** Let D be a bounded smooth domain, and define

$$A_{a,c}u := -\sum_{i,j=1}^{d} \partial_i\big(a_{ij}(x)\partial_j u\big) + c(x)u$$

with homogeneous conormal Neumann boundary condition

$$(a\nabla u) \cdot \nu = 0 \quad \text{on } \partial D.$$

Let $\mathcal{C}$ be a class of real smooth coefficients (a, c) on $\overline{D}$ such that $a = a^\top$ and

$$\lambda|\xi|^2 \leqslant \sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \leqslant \Lambda|\xi|^2, \quad 0 \leqslant c(x) \leqslant C_0$$

for any $x \in \overline{D}$ and $\xi \in \mathbb{R}^d$, where $0 < \lambda \leqslant \Lambda$ and $C_0 > 0$ are fixed. Fix a locally Lipschitz function $F : \mathbb{R} \to \mathbb{R}$ with $F(0) = 0$ and $\mathrm{Lip}_{[-M,M]}(F) \leqslant L$, and set

$$\Omega_{\mathrm{eq}} := \{(A_{a,c}, F) : (a, c) \in \mathcal{C}\}.$$

We use $C(\overline{D})$ as the continuous function space. The associated semigroups are strongly continuous on this space and are L^∞-contractive. Thus, the assumptions of Proposition B.1 hold with $C_{\mathcal{A}}(T) = 1$ and $\mathcal{F} = \{F\}$. Hence, Assumption 3.1 holds with $C_S = 1$.

Here, the unknown equation specification is given by the coefficients (a, c).

Therefore, these classes satisfy Assumption 3.1 and hence, the equation-class assumption of Theorem 5.1.

B.5 Scope and possible extensions

Proposition B.1 gives uniform Picard contraction on a sufficiently short time interval under common semigroup and local Lipschitz bounds, together with the required continuity assumptions. The nonlinearity itself need not be small. More generally, in many scaling-subcritical settings, the contraction factor can be made small by shortening the time interval. Uniform contraction over an equation class, however, requires uniform bounds on the constants in these estimates.

The present analysis uses continuous solution spaces equipped with the $L^\infty_{t,x}$ norm. This setting is natural for many semilinear parabolic equations [Henry, 1981, Pazy, 1983]. For nonlinear wave

equations or the Navier–Stokes equations, mixed spaces $L^r_t L^s_x$ or Sobolev spaces may be more suitable [Giga and Miyakawa, 1985, Cannone, 2004, Cazenave and Haraux, 1998, Tao, 2006]. Extending the framework to these spaces requires contraction, stability, and complexity estimates for the corresponding Picard class, together with suitable observation and readout maps.

The criterion directly covers reaction terms $F = F(u)$ that are locally Lipschitz. Derivative-dependent terms, such as $|\nabla u|^p$, require additional regularity and different estimates for the Duhamel map. For viscous Hamilton–Jacobi equations, related fixed point arguments in suitable subcritical spaces are available; see Ben-Artzi et al. [2002]. Sublinear terms $|u|^p$ with $0 < p < 1$ and logarithmic terms $u \log |u|$ are not locally Lipschitz at zero. Their well-posedness and contraction properties therefore require separate analysis and do not follow from Proposition B.1. Depending on the equation, other methods, such as monotonicity or regularization, may be needed [Cazenave and Haraux, 1998, Pao, 1992, Crandall and Liggett, 1971, Barbu, 2010].

Finally, short-time Picard contraction alone does not guarantee stable long-time prediction. Section 6 analyzes repeated use of a fixed local model under additional conditions on rollout admissibility and local errors. Extensions to other iterative methods are discussed in Appendix G.

C Picard Truncation Error Bounds: Proofs of Lemmas 4.1 and 4.2

In this appendix we prove Lemmas 4.1 and 4.2. We write $d(u, v) := \|u - v\|_{L^\infty(Q_T)}$ for $u, v \in U_M$.

Lemma C.1 (Lemma 4.1)**.** *Suppose that Assumption 3.1 holds. Then, for any $\omega \in \Omega_{\mathrm{eq}}$ and $\ell \in \mathbb{N}$, we have*

$$\sup_{v \in U_{0,R}} \|b_{\omega,\ell}(v, \cdot) - \mathcal{G}_\omega[v]\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta}. \tag{26}$$

Consequently,

$$\inf_{b \in \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} L[b] \leqslant \frac{M^2\delta^{2\ell}}{(1-\delta)^2}. \tag{27}$$

Proof of Lemma 4.1. Fix $\omega \in \Omega_{\mathrm{eq}}$ and $v \in U_{0,R}$. Set $u^{(0)} := 0$ and

$$u^{(\ell+1)} := \mathcal{T}_{\omega,v}[u^{(\ell)}], \quad \ell = 0, 1, 2, \cdots.$$

Then, $b_{\omega,\ell}(v, \cdot) = u^{(\ell)} = \mathcal{T}^{[\ell]}_{\omega,v}[0]$ by the definition of $b_{\omega,\ell}$. Let $u := \mathcal{G}_\omega[v]$ be the unique fixed point of $\mathcal{T}_{\omega,v}$. Since $\mathcal{T}_{\omega,v}$ is δ-contractive on U_M, we have

$$d(u^{(r+1)}, u^{(r)}) \leqslant \delta^r d(u^{(1)}, u^{(0)}), \quad r = 0, 1, 2, \cdots.$$

Therefore, for $N > \ell$, we estimate

$$d(u^{(N)}, u^{(\ell)}) \leqslant \sum_{r=\ell}^{N-1} d(u^{(r+1)}, u^{(r)}) \leqslant \sum_{r=\ell}^{N-1} \delta^r d(u^{(1)}, u^{(0)}) \leqslant \frac{\delta^\ell}{1-\delta} d(u^{(1)}, u^{(0)}).$$

Letting $N \to \infty$ and using $u^{(N)} \to u$ in $L^\infty(Q_T)$, we obtain

$$d(u^{(\ell)}, u) \leqslant \frac{\delta^\ell}{1-\delta} d(u^{(1)}, u^{(0)}).$$

Since $u^{(1)} = \mathcal{T}_{\omega,v}[0] \in U_M$ and $u^{(0)} = 0$, we have $d(u^{(1)}, u^{(0)}) = \|u^{(1)}\|_{L^\infty(Q_T)} \leqslant M$. Hence

$$\|b_{\omega,\ell}(v, \cdot) - \mathcal{G}_\omega[v]\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta}.$$

Taking the supremum over $v \in U_{0,R}$ gives (26).

We next prove the bound (27) for the oracle term. Let $\omega_* \in \Omega_{\mathrm{eq}}$ be such that $\mathcal{G}_* = \mathcal{G}_{\omega_*}$. Since $b_{\omega_*,\ell} \in \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}$, we have

$$\inf_{b \in \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} L[b] \leqslant L[b_{\omega_*,\ell}] = \mathbb{E}_{u_0 \sim \Pi, z \sim \mu}\left[\left|b_{\omega_*,\ell}(u_0, z) - \mathcal{G}_{\omega_*}[u_0](z)\right|^2\right] \leqslant \left(\frac{M\delta^\ell}{1-\delta}\right)^2.$$

Thus, (27) is shown. The proof of Lemma C.1 is finished. □

Lemma C.2 (Lemma 4.2)**.** *Suppose that Assumption 3.1 holds. Then, for any $\omega \in \Omega_{\mathrm{eq}}$, $E \in \mathcal{E}_m$, $u_0 \in U_0$ and $\ell \in \mathbb{N}$, we have*

$$\|b_{\omega,E,\ell}(P_m u_0, \cdot) - \mathcal{G}_\omega[u_0]\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta} + \frac{C_S}{1-\delta}\|E(P_m u_0) - u_0\|_{L^\infty(D)}. \tag{28}$$

Consequently,

$$\inf_{b \in \mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}}} L_m[b] \leqslant \frac{2M^2\delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2}{(1-\delta)^2}\varepsilon_{\mathrm{rec}}^2(m). \tag{29}$$

Proof of Lemma 4.2. Fix $\omega \in \Omega_{\mathrm{eq}}$, $E \in \mathcal{E}_m$, and $u_0 \in U_0$. Put

$$\widetilde{u}_0 := E(P_m u_0).$$

Since $E : \mathbb{R}^m \to U_0$ and $U_0 \subset U_{0,R}$, we have $\widetilde{u}_0 \in U_{0,R}$. By the definition of the finite-observation Picard model,

$$b_{\omega,E,\ell}(P_m u_0, \cdot) = \mathcal{T}_{\omega,\widetilde{u}_0}^{[\ell]}[0].$$

Using Lemma C.1 with initial data $\widetilde{u}_0$, we obtain

$$\|b_{\omega,E,\ell}(P_m u_0, \cdot) - \mathcal{G}_\omega[\widetilde{u}_0]\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta}.$$

On the other hand, by (4) (i.e. the continuous dependence on initial data in Proposition 3.2), we have

$$\|\mathcal{G}_\omega[\widetilde{u}_0] - \mathcal{G}_\omega[u_0]\|_{L^\infty(Q_T)} \leqslant \frac{C_S}{1-\delta}\|\widetilde{u}_0 - u_0\|_{L^\infty(D)}.$$

Therefore, by the triangle inequality, we obtain (28):

$$\begin{aligned}
&\|b_{\omega,E,\ell}(P_m u_0, \cdot) - \mathcal{G}_\omega[u_0]\|_{L^\infty(Q_T)} \\
&\quad\leqslant \|b_{\omega,E,\ell}(P_m u_0, \cdot) - \mathcal{G}_\omega[\widetilde{u}_0]\|_{L^\infty(Q_T)} + \|\mathcal{G}_\omega[\widetilde{u}_0] - \mathcal{G}_\omega[u_0]\|_{L^\infty(Q_T)} \\
&\quad\leqslant \frac{M\delta^\ell}{1-\delta} + \frac{C_S}{1-\delta}\|E(P_m u_0) - u_0\|_{L^\infty(D)}.
\end{aligned}$$

We now prove the corresponding oracle bound (29). Let $\omega_* \in \Omega_{\mathrm{eq}}$ be such that $\mathcal{G}_* = \mathcal{G}_{\omega_*}$. For each $E \in \mathcal{E}_m$, the preceding estimate gives

$$\left|b_{\omega_*,E,\ell}(P_m u_0, z) - \mathcal{G}_*[u_0](z)\right|^2 \leqslant 2\left(\frac{M\delta^\ell}{1-\delta}\right)^2 + 2\left(\frac{C_S}{1-\delta}\right)^2 \|E(P_m u_0) - u_0\|_{L^\infty(D)}^2$$

for any $u_0 \in U_0$ and $z \in Q_T$. Taking expectation with respect to $u_0 \sim \Pi$ and $z \sim \mu$, we obtain

$$L_m[b_{\omega_*,E,\ell}] \leqslant \frac{2M^2\delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2}{(1-\delta)^2}\mathbb{E}_{u_0\sim\Pi}\left[\|E(P_m u_0) - u_0\|_{L^\infty(D)}^2\right].$$

Since $b_{\omega_*,E,\ell} \in \mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}}$, it follows that

$$\begin{aligned}
\inf_{b \in \mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}}} L_m[b] &\leqslant \inf_{E \in \mathcal{E}_m} L_m[b_{\omega_*,E,\ell}] \\
&\leqslant \frac{2M^2\delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2}{(1-\delta)^2} \inf_{E \in \mathcal{E}_m} \mathbb{E}_{u_0\sim\Pi}\left[\|E(P_m u_0) - u_0\|_{L^\infty(D)}^2\right].
\end{aligned}$$

By the definition of $\varepsilon_{\mathrm{rec}}^2(m)$, this yields (29). The proof of Lemma C.2 is finished. □

D Generalization Error Bounds: Proofs of Theorem 5.1 and Corollary 5.3

In this appendix we prove Theorem 5.1 and Corollary 5.3. We first record the uniform deviation estimate used in the proof. The statement is formulated at the trajectory-block level so that it applies both to the full-information case and to the finite-observation case.

Remark D.1 (Uniform deviation estimate for the clipped loss). Let $\mathcal{X}$ be an input space and let $\mathcal{B} \subset \{b : \mathcal{X} \times Q_T \to [-M, M]\}$. Let W be a latent trajectory variable with law ν, let $A(W) \in \mathcal{X}$ be the model input, and let $y(W, z) \in [-M, M]$ be the target value at the query point z. For independent blocks $(W_i, z_{i,1}, \dots, z_{i,q}) \sim \nu \otimes \mu^{\otimes q}$, set $a_i = A(W_i)$ and $y_{i,j} = y(W_i, z_{i,j})$. For the trajectory-level sample $\mathcal{S}^{\mathcal{X}}_{n,q} = \{(a_i, \{z_{i,j}\}_{j=1}^q)\}_{i=1}^n$, define the pointwise empirical Rademacher complexity by

$$\widehat{\mathfrak{R}}^{\mathrm{pt}}_{\mathcal{S}^{\mathcal{X}}_{n,q}}(\mathcal{B}) := \mathbb{E}_\sigma \left[\sup_{b \in \mathcal{B}} \frac{1}{nq} \sum_{i=1}^n \sum_{j=1}^q \sigma_{i,j} b(a_i, z_{i,j}) \right],$$

where $\{\sigma_{i,j} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant q\}$ are independent Rademacher variables. As in Section 5, we use the trajectory-normalized version

$$\widehat{\mathfrak{R}}_{\mathcal{S}^{\mathcal{X}}_{n,q}}(\mathcal{B}) := \sqrt{q}\, \widehat{\mathfrak{R}}^{\mathrm{pt}}_{\mathcal{S}^{\mathcal{X}}_{n,q}}(\mathcal{B}) = \mathbb{E}_\sigma \left[\sup_{b \in \mathcal{B}} \frac{1}{n\sqrt{q}} \sum_{i=1}^n \sum_{j=1}^q \sigma_{i,j} b(a_i, z_{i,j}) \right].$$

Define

$$L_{\mathcal{X}}[b] := \mathbb{E}_{W \sim \nu, z \sim \mu} \ell_{\Lambda_{\mathrm{clip}}}\big(b(A(W), z), y(W, z)\big)$$

and

$$\widehat{L}_{\mathcal{X},n,q}[b] := \frac{1}{n} \sum_{i=1}^n \frac{1}{q} \sum_{j=1}^q \ell_{\Lambda_{\mathrm{clip}}}\big(b(a_i, z_{i,j}), y_{i,j}\big).$$

Then there exists a universal constant $C > 0$ such that, with probability at least $1 - \zeta$,

$$\sup_{b \in \mathcal{B}} \left| L_{\mathcal{X}}[b] - \widehat{L}_{\mathcal{X},n,q}[b] \right| \leqslant CM \widehat{\mathfrak{R}}_{\mathcal{S}^{\mathcal{X}}_{n,q}}(\mathcal{B}) + CM^2 \sqrt{\frac{\log(2/\zeta)}{n}}.$$

Indeed, symmetrization is applied to the n independent trajectory blocks. Conditional on the sample, define

$$\phi_i(r_1, \dots, r_q) := \frac{1}{q} \sum_{j=1}^q \ell_{\Lambda_{\mathrm{clip}}}(r_j, y_{i,j}).$$

Since the clipped squared loss is CM-Lipschitz in the scalar prediction on $[-M, M]$, each ϕ_i is $CM/\sqrt{q}$-Lipschitz with respect to the Euclidean norm on $\mathbb{R}^q$. Maurer's vector contraction inequality [Maurer, 2016] therefore controls the Rademacher complexity of the clipped loss class by $CM\widehat{\mathfrak{R}}_{\mathcal{S}^{\mathcal{X}}_{n,q}}(\mathcal{B})$. The last term follows from bounded-difference concentration for the block losses, which are bounded by CM^2.

Proof of Theorem 5.1. We first prove the full-information bound. Put

$$\mathcal{B}_\ell := \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}.$$

Since all predictors are clipped to $[-M, M]$ and $\mathcal{G}_*[u_0] \in U_M$, the target values $\mathcal{G}_*[u_0](z)$ also belong to $[-M, M]$. Hence the clipped squared loss is bounded by $4M^2$ and is $4M$-Lipschitz in the first variable.

Apply Remark D.1 with $W = u_0$, $A(W) = u_0$, and $y(W, z) = \mathcal{G}_*(u_0)(z)$. Then, with probability at least $1 - \zeta$, we have

$$\sup_{b \in \mathcal{B}_\ell} \left| L[b] - \widehat{L}_{n,q}[b] \right| \leqslant CM \widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} + CM^2 \sqrt{\frac{\log(2/\zeta)}{n}}.$$

Denote the right-hand side by Δ_ℓ. Thus, on this event,

$$L[b] \leqslant \widehat{L}_{n,q}[b] + \Delta_\ell, \qquad \widehat{L}_{n,q}[b] \leqslant L[b] + \Delta_\ell$$

for every $b \in \mathcal{B}_\ell$.

Let $\widehat{b}_\ell \in \operatorname{argmin}_{b \in \mathcal{B}_\ell} \widehat{L}_{n,q}[b]$ and $\widehat{\Gamma}_\ell := \mathrm{embed}(\widehat{b}_\ell)$. By the definition of the implementation error,

$$L[\widehat{\Gamma}_\ell] = L[\mathrm{embed}(\widehat{b}_\ell)] \leqslant L[\widehat{b}_\ell] + \varepsilon_{\mathrm{imp}}.$$

Using the uniform deviation estimate and the empirical minimality of $\widehat{b}_\ell$, we obtain, for every $b \in \mathcal{B}_\ell$,

$$L[\widehat{b}_\ell] \leqslant \widehat{L}_{n,q}[\widehat{b}_\ell] + \Delta_\ell \leqslant \widehat{L}_{n,q}[b] + \Delta_\ell \leqslant L[b] + 2\Delta_\ell.$$

Taking the infimum over $b \in \mathcal{B}_\ell$ gives

$$L[\widehat{b}_\ell] \leqslant \inf_{b \in \mathcal{B}_\ell} L[b] + 2\Delta_\ell.$$

Therefore,

$$L[\widehat{\Gamma}_\ell] \leqslant \varepsilon_{\mathrm{imp}} + \inf_{b \in \mathcal{B}_\ell} L[b] + 2\Delta_\ell.$$

By Lemma 4.1,

$$\inf_{b \in \mathcal{B}_\ell} L[b] \leqslant \frac{M^2 \delta^{2\ell}}{(1-\delta)^2}.$$

Absorbing the factor 2 into the universal constant C, we obtain

$$L[\widehat{\Gamma}_\ell] \leqslant \varepsilon_{\mathrm{imp}} + \frac{M^2 \delta^{2\ell}}{(1-\delta)^2} + CM\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}.$$

This proves part (i).

We next prove the finite-observation bound. Put

$$\mathcal{B}_{m,\ell} := \mathcal{B}_{m,\ell}^{\Omega_{\mathrm{eq}}}.$$

Apply the same uniform deviation estimate with $W = u_0$, $A(W) = P_m u_0$, and $y(W, z) = \mathcal{G}_*(u_0)(z)$ to the input space $P_m(U_0)$ and the class $\mathcal{B}_{m,\ell}$. Then, with probability at least $1 - \zeta$,

$$\sup_{b \in \mathcal{B}_{m,\ell}} \left| L_m[b] - \widehat{L}_{m,n,q}[b] \right| \leqslant CM\widehat{\mathfrak{R}}_m^{\Omega_{\mathrm{eq}}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}.$$

Denote the right-hand side by $\Delta_{m,\ell}$. Let $\widehat{b}_{m,\ell} \in \operatorname{argmin}_{b \in \mathcal{B}_{m,\ell}} \widehat{L}_{m,n,q}[b]$ and $\widehat{\Gamma}_{m,\ell} := \mathrm{embed}_m(\widehat{b}_{m,\ell})$. Then, by the definition of $\varepsilon_{\mathrm{imp}}^m$,

$$L_m[\widehat{\Gamma}_{m,\ell}] = L_m[\mathrm{embed}_m(\widehat{b}_{m,\ell})] \leqslant L_m[\widehat{b}_{m,\ell}] + \varepsilon_{\mathrm{imp}}^m.$$

As before, the empirical minimality of $\widehat{b}_{m,\ell}$ implies that, for every $b \in \mathcal{B}_{m,\ell}$,

$$L_m[\widehat{b}_{m,\ell}] \leqslant \widehat{L}_{m,n,q}[\widehat{b}_{m,\ell}] + \Delta_{m,\ell} \leqslant \widehat{L}_{m,n,q}[b] + \Delta_{m,\ell} \leqslant L_m[b] + 2\Delta_{m,\ell}.$$

Taking the infimum over $b \in \mathcal{B}_{m,\ell}$, we get

$$L_m[\widehat{b}_{m,\ell}] \leqslant \inf_{b \in \mathcal{B}_{m,\ell}} L_m[b] + 2\Delta_{m,\ell}.$$

Thus

$$L_m[\widehat{\Gamma}_{m,\ell}] \leqslant \varepsilon_{\mathrm{imp}}^m + \inf_{b \in \mathcal{B}_{m,\ell}} L_m[b] + 2\Delta_{m,\ell}.$$

By Lemma 4.2,

$$\inf_{b \in \mathcal{B}_{m,\ell}} L_m[b] \leqslant \frac{2M^2 \delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2}{(1-\delta)^2} \varepsilon_{\mathrm{rec}}^2(m).$$

Therefore,

$$L_m[\widehat{\Gamma}_{m,\ell}] \leqslant \varepsilon_{\mathrm{imp}}^m + \frac{2M^2 \delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2}{(1-\delta)^2} \varepsilon_{\mathrm{rec}}^2(m) + CM\widehat{\mathfrak{R}}_m^{\Omega_{\mathrm{eq}}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}.$$

This proves part (ii).

If one wants the two estimates to hold simultaneously, one applies the above two deviation estimates with confidence parameter $\zeta/2$ and absorbs the resulting harmless factor into the universal constant C. □

Proof of Corollary 5.3. We first consider the full-information case. By the entropy bound stated in Section 5, we have

$$\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} \leqslant \frac{C}{\sqrt{n}} \mathcal{I}_\Omega,$$

where

$$\mathcal{I}_\Omega := \int_0^{2M} \sqrt{\mathcal{H}_\Omega(c(1-\delta)\varepsilon)}\, d\varepsilon < \infty.$$

Let

$$\ell_n := \left\lceil \frac{\log n}{4|\log \delta|} \right\rceil.$$

Since $0 < \delta < 1$, we have $|\log \delta| = -\log \delta$. Hence

$$\delta^{2\ell_n} = \exp(-2\ell_n |\log \delta|) \leqslant \exp\left(-\frac{1}{2}\log n\right) = n^{-1/2}.$$

Therefore,

$$\frac{M^2 \delta^{2\ell_n}}{(1-\delta)^2} \leqslant C \frac{M^2}{\sqrt{n}},$$

where the constant C may depend on δ, but not on n. Substituting these estimates into Theorem 5.1 gives

$$L[\widehat{\Gamma}_{\ell_n}] \leqslant \varepsilon_{\mathrm{imp}} + C\left(\frac{M^2}{\sqrt{n}} + \frac{M\mathcal{I}_\Omega}{\sqrt{n}} + M^2\sqrt{\frac{\log(2/\zeta)}{n}}\right).$$

This proves the full-information estimate.

We next consider the finite-observation case. Assume that

$$\varepsilon_{\mathrm{rec}}^2(m) \leqslant C_{\mathrm{rec}}^2 m^{-2\beta}$$

and

$$\mathcal{I}_{\Omega,m} \leqslant C_I m^\alpha.$$

Using the entropy estimate $\widehat{\mathfrak{R}}_m^{\Omega_{\mathrm{eq}}} \leqslant C\mathcal{I}_{\Omega,m}/\sqrt{n}$, Theorem 5.1 gives

$$\begin{aligned} L_m[\widehat{\Gamma}_{m,\ell}] \leqslant \varepsilon_{\mathrm{imp}}^m + \frac{2M^2\delta^{2\ell}}{(1-\delta)^2} + \frac{2C_S^2 C_{\mathrm{rec}}^2}{(1-\delta)^2} m^{-2\beta} \\ + CMC_I\frac{m^\alpha}{\sqrt{n}} + CM^2\sqrt{\frac{\log(1/\zeta)}{n}}. \end{aligned}$$

We now take $\ell = \ell_n$ and $m = m_n$, where

$$m_n \asymp n^{1/(2(2\beta+\alpha))}.$$

The choice of ℓ_n gives

$$\delta^{2\ell_n} \leqslant n^{-1/2}.$$

Moreover, by the choice of m_n,

$$m_n^{-2\beta} \asymp n^{-\beta/(2\beta+\alpha)}$$

and

$$\frac{m_n^\alpha}{\sqrt{n}} \asymp n^{\alpha/(2(2\beta+\alpha))} n^{-1/2} = n^{-\beta/(2\beta+\alpha)}.$$

Since $\alpha \geqslant 0$, we have $\beta/(2\beta+\alpha) \leqslant 1/2$, and hence $n^{-1/2} \leqslant n^{-\beta/(2\beta+\alpha)}$ for $n \geqslant 1$. Therefore the Picard truncation term is also bounded by a constant multiple of $n^{-\beta/(2\beta+\alpha)}$. Consequently,

$$L_{m_n}[\widehat{\Gamma}_{m_n,\ell_n}] \leqslant \varepsilon_{\mathrm{imp}}^{m_n} + C\left(n^{-\beta/(2\beta+\alpha)} + M^2\sqrt{\frac{\log(2/\zeta)}{n}}\right),$$

where C may depend on $M, C_S, \delta, C_{\mathrm{rec}}, C_I$, but not on n. This proves the finite-observation estimate.

For quasi-uniform sensors and stable reconstruction on $U_0 \subset H^{s_0}(D)$ with $s_0 > d/2$, the usual Sobolev reconstruction rate gives

$$\beta = \frac{s_0 - d/2}{d}.$$

This completes the proof of Corollary 5.3. □

E Long-Time Prediction: Proofs of Theorem 6.1

In this appendix, we consider long-time prediction based on the learned local model.

The idea is based on the standard argument for extending local-in-time solutions. First, let us briefly recall the local-in-time solution theory. Suppose that $\Omega_{\mathrm{eq}} := \mathcal{A} \times \mathcal{F}$ is an admissible equation class satisfying the assumptions of Proposition B.1. Then, for any R, M with $0 < R < M$ and any $\delta \in (0,1)$, we can choose $T = T(R, M, \delta) > 0$ sufficiently small so that (17) holds. Hence, by Proposition B.1, Assumption 3.1 holds for this choice of R, M, δ and T. Consequently, the local well-posedness result in Proposition 3.2 follows from Banach's fixed point theorem. In other words, for each $\omega \in \Omega_{\mathrm{eq}}$, there exists a unique solution operator $\mathcal{G}_\omega$ which maps an initial datum $u_0 \in U_{0,R}$ to the corresponding solution u on $[0,T]$ of (P').

If the terminal value $u(T)$ belongs to $U_{0,R}$, that is, $\|u(T)\|_{L^\infty(D)} \leqslant R$, then $u(T)$ can be used as a new initial data. Thus, the solution can be extended to $[T, 2T]$ by applying the same local solution operator $\mathcal{G}_\omega$ again. More precisely, after the time shift $s = t - T$, the solution on $[T, 2T]$ is given by

$$u(T+s) = \mathcal{G}_\omega[u(T)](s), \qquad 0 \leqslant s \leqslant T.$$

Equivalently, if we define the terminal map $\psi(u) := u(T)$ and $\Phi_\omega := \psi \circ \mathcal{G}_\omega$, then the second-step initial data is $\Phi_\omega(u_0)$, and the solution on the second time block is

$$\mathcal{G}_\omega[\Phi_\omega(u_0)].$$

Repeating this procedure as long as the terminal values remain in $U_{0,R}$, we can extend the solution successively over longer time intervals. The idea of the long-time prediction result is to use this standard continuation mechanism at the level of the learned local solution operator.

This idea has three main advantages. First, the same local solution operator is used at every time block. Therefore, once the solution operator has been learned well on the first time interval $[0,T]$, no additional learning is needed for later time intervals. Long-time prediction can be obtained by repeatedly using the short-time learned model.

Second, it is enough to check the size of the terminal value. As long as the terminal value remains in $U_{0,R}$, or equivalently its $L^\infty(D)$ norm remains bounded by R, the same local solution operator can be applied again. Thus, one can predict the solution with a theoretical guarantee, although the accumulated error of the short-time learned model must be taken into account.

Third, for PDEs whose solutions are uniformly bounded, the exact solution does not exceed the threshold R if R is chosen appropriately. Moreover, one can often show that a short-time learned model, such as a neural operator, can approximate the local solution operator well while keeping the output uniformly bounded by R (see Remark F.3 and Furuya et al. [2025]). In such a case, the rollout remains in the admissible state set, and the same local error estimate can be iterated over successive time blocks.

E.1 Formulation and Result

Assume that every trajectory considered below admits a terminal trace. In our present setting, this is automatic, since the solution space is defined as $U_M := \{u \in C([0,T]; C(D)) : \|u\|_{L^\infty(Q_T)} \leqslant M\}$. Define $\psi : U_M \to L^\infty(D)$ by $\psi(u) := u(\cdot, T)$, and assume that

$$\|\psi(u) - \psi(v)\|_{L^\infty(D)} \leqslant L_\psi \|u - v\|_{L^\infty(Q_T)}.$$

Let $\Phi_* := \psi \circ \mathcal{G}_*$. For $\kappa \in \mathbb{N}$, define $U_0^{[\kappa]} := \{u_0 \in U_0 : \Phi_*^j(u_0) \in U_{0,R} \text{ for } j = 0, \dots, \kappa\}$. This condition ensures that the same local solution theory can be applied on each time block. We assume $\Pi(U_0^{[\kappa]}) > 0$ and write $\Pi_\kappa := \Pi(\cdot \mid U_0^{[\kappa]})$.

For $u_0 \in U_0^{[\kappa]}$, define the exact states by $v_0 = u_0$ and $v_{j+1} = \Phi_*(v_j)$. The exact solution on the j-th block is $\mathcal{G}_*(v_j)$ after the time shift $t = jT + s$, $s \in [0,T]$. Next, define the approximate transition by $\widehat{\Phi}_\ell := \mathcal{P}_R \circ \psi \circ \widehat{\Gamma}_\ell$, where $\mathcal{P}_R : C(D) \to U_{0,R}$ is the pointwise clipping map. The approximate states are $\widehat{v}_0 = u_0$ and $\widehat{v}_{j+1} = \widehat{\Phi}_\ell(\widehat{v}_j)$. Thus, the model prediction on the j-th block is $\widehat{\Gamma}_\ell[\widehat{v}_j]$. For $j = 0, \dots, \kappa - 1$, define the block-wise risk by

$$L_j^{\mathrm{block}}[\widehat{\Gamma}_\ell] := \mathbb{E}_{u_0 \sim \Pi_\kappa,\, z \sim \mu}\left[\left|\widehat{\Gamma}_\ell[\widehat{v}_j](z) - \mathcal{G}_*[v_j](z)\right|^2\right].$$

Let $\mathcal{V}_\kappa(\widehat{\Gamma}_\ell) := \{v_j(u_0), \widehat{v}_j(u_0) : u_0 \in U_0^{[\kappa]},\ j = 0, \dots, \kappa\} \subset U_{0,R}$, and define the rollout-relevant local error by

$$\varepsilon_{\mathrm{loc},\ell}^{[\kappa]} := \sup_{v \in \mathcal{V}_\kappa(\widehat{\Gamma}_\ell)} \left\| \widehat{\Gamma}_\ell[v] - \mathcal{G}_*[v] \right\|_{L^\infty(Q_T)}. \tag{30}$$

Theorem E.1 (Theorem 6.1, Block-wise long-time error propagation)**.** *Assume Assumption 3.1 and the terminal trace condition above. Then, for any $j = 0, \dots, \kappa - 1$,*

$$L_j^{\mathrm{block}}[\widehat{\Gamma}_\ell] \leqslant \left(\sum_{r=0}^{j} \left(L_\psi \frac{C_S}{1-\delta} \right)^r \right)^2 \left(\varepsilon_{\mathrm{loc},\ell}^{[\kappa]} \right)^2.$$

This result shows that long-time prediction is obtained by deterministic propagation of the local error. Since no additional empirical risk minimization is performed after $\widehat{\Gamma}_\ell$ has been learned, the rollout step does not introduce a new Rademacher complexity term. The block index j enters only through the stability factor in Theorem E.1. The proof iterates the local stability estimate (4) along the exact and approximate trajectories.

E.2 Proof of Theorem E.1

For simplicity, we write

$$L_G := \frac{C_S}{1-\delta}, \quad A_G := L_\psi L_G = L_\psi \frac{C_S}{1-\delta}, \quad \varepsilon := \varepsilon_{\mathrm{loc},\ell}^{[\kappa]}.$$

Recall that

$$v_{j+1} = \Phi_*(v_j) = \psi\big(\mathcal{G}_*[v_j]\big), \quad \widehat{v}_{j+1} = \widehat{\Phi}_\ell(\widehat{v}_j) = \mathcal{P}_R \psi\big(\widehat{\Gamma}_\ell[\widehat{v}_j]\big).$$

Since $v_j, \widehat{v}_j \in \mathcal{V}_\kappa(\widehat{\Gamma}_\ell)$ for $j = 0, \dots, \kappa$, the definition of ε gives

$$\left\| \widehat{\Gamma}_\ell[w] - \mathcal{G}_*[w] \right\|_{L^\infty(Q_T)} \leqslant \varepsilon$$

for any $w \in \{v_j, \widehat{v}_j : j = 0, \dots, \kappa\}$. We also use that the pointwise clipping map $\mathcal{P}_R$ is nonexpansive in $L^\infty(D)$, namely

$$\|\mathcal{P}_R f - \mathcal{P}_R g\|_{L^\infty(D)} \leqslant \|f - g\|_{L^\infty(D)}.$$

Moreover, since $v_{j+1} \in U_{0,R}$, we have $\mathcal{P}_R v_{j+1} = v_{j+1}$. Define the state error

$$e_j := \|\widehat{v}_j - v_j\|_{L^\infty(D)}.$$

Then $e_0 = 0$. We first prove the recursive estimate

$$e_{j+1} \leqslant L_\psi \varepsilon + A_G e_j, \quad j = 0, \dots, \kappa - 1.$$

In fact, using $v_{j+1} = \psi(\mathcal{G}_*[v_j])$ and $\mathcal{P}_R v_{j+1} = v_{j+1}$, we have

$$\begin{aligned} e_{j+1} &= \|\widehat{v}_{j+1} - v_{j+1}\|_{L^\infty(D)} = \left\| \mathcal{P}_R \psi\big(\widehat{\Gamma}_\ell[\widehat{v}_j]\big) - \mathcal{P}_R \psi\big(\mathcal{G}_*[v_j]\big) \right\|_{L^\infty(D)} \\ &\leqslant \left\| \psi\big(\widehat{\Gamma}_\ell[\widehat{v}_j]\big) - \psi\big(\mathcal{G}_*[v_j]\big) \right\|_{L^\infty(D)} \leqslant L_\psi \left\| \widehat{\Gamma}_\ell[\widehat{v}_j] - \mathcal{G}_*[v_j] \right\|_{L^\infty(Q_T)}. \end{aligned}$$

By adding and subtracting $\mathcal{G}_*[\widehat{v}_j]$, we obtain

$$\left\| \widehat{\Gamma}_\ell[\widehat{v}_j] - \mathcal{G}_*[v_j] \right\|_{L^\infty(Q_T)} \leqslant \left\| \widehat{\Gamma}_\ell[\widehat{v}_j] - \mathcal{G}_*[\widehat{v}_j] \right\|_{L^\infty(Q_T)} + \|\mathcal{G}_*[\widehat{v}_j] - \mathcal{G}_*[v_j]\|_{L^\infty(Q_T)}.$$

The first term is bounded by ε by the definition of $\varepsilon_{\mathrm{loc},\ell}^{[\kappa]}$. For the second term, since $\mathcal{G}_* = \mathcal{G}_{\omega_*}$ and both v_j and $\widehat{v}_j$ belong to $U_{0,R}$, the estimate (4) gives

$$\|\mathcal{G}_*[\widehat{v}_j] - \mathcal{G}_*[v_j]\|_{L^\infty(Q_T)} \leqslant L_G \|\widehat{v}_j - v_j\|_{L^\infty(D)} = L_G e_j.$$

Therefore,

$$e_{j+1} \leqslant L_\psi(\varepsilon + L_G e_j) = L_\psi \varepsilon + A_G e_j.$$

This proves the recursive estimate. We now solve this recursion. Since $e_0 = 0$, induction gives

$$e_j \leqslant L_\psi \varepsilon \sum_{r=0}^{j-1} A_G^r, \quad j = 1, \dots, \kappa.$$

For $j = 0$, this estimate is understood as $e_0 = 0$.

Next, for the prediction error on the j-th block, we estimate

$$\begin{aligned} \left\| \widehat{\Gamma}_\ell[\widehat{v}_j] - \mathcal{G}_*[v_j] \right\|_{L^\infty(Q_T)} &\leqslant \left\| \widehat{\Gamma}_\ell[\widehat{v}_j] - \mathcal{G}_*[\widehat{v}_j] \right\|_{L^\infty(Q_T)} + \| \mathcal{G}_*[\widehat{v}_j] - \mathcal{G}_*[v_j] \|_{L^\infty(Q_T)} \\ &\leqslant \varepsilon + L_G e_j. \end{aligned}$$

Using the bound on e_j, we obtain

$$\left\| \widehat{\Gamma}_\ell[\widehat{v}_j] - \mathcal{G}_*[v_j] \right\|_{L^\infty(Q_T)} \leqslant \varepsilon + L_G L_\psi \varepsilon \sum_{r=0}^{j-1} A_G^r = \varepsilon \left(1 + A_G \sum_{r=0}^{j-1} A_G^r \right) = \varepsilon \sum_{r=0}^{j} A_G^r.$$

Hence, for any $z \in Q_T$,

$$\left| \widehat{\Gamma}_\ell[\widehat{v}_j](z) - \mathcal{G}_*[v_j](z) \right|^2 \leqslant \left(\sum_{r=0}^{j} A_G^r \right)^2 \varepsilon^2.$$

Taking expectation with respect to $u_0 \sim \Pi_\kappa$ and $z \sim \mu$, we obtain

$$L_j^{\mathrm{block}}[\widehat{\Gamma}_\ell] \leqslant \left(\sum_{r=0}^{j} A_G^r \right)^2 \varepsilon^2.$$

Substituting $A_G = L_\psi C_S/(1-\delta)$ and $\varepsilon = \varepsilon_{\mathrm{loc},\ell}^{[\kappa]}$, we conclude Theorem E.1.

F Nonlinear Heat Equations on the Torus and Picard-Type FNOs

F.1 Nonlinear Heat Equations on the Torus and the Operator Learning Setting

We specialize the abstract Picard-type operator learning framework of Sections 3–6 to nonlinear heat equations on the torus. Let $\mathbb{T}^d := \mathbb{R}^d/(2\pi\mathbb{Z})^d$ and $Q_T := \mathbb{T}^d \times (0,T)$. We consider

$$\partial_t u - \Delta_{\mathbb{T}^d} u = F(u), \quad u(0) = u_0. \tag{31}$$

The linear part is fixed and the unknown equation specification is the nonlinearity. For $R, M, L > 0$, set

$$\mathcal{F}_{M,L} := \left\{ F : [-M, M] \to \mathbb{R} : F(0) = 0,\ \mathrm{Lip}_{[-M,M]}(F) \leqslant L \right\},$$

and define the admissible equation class by

$$\Omega_{\mathrm{eq}} := \{ \omega_F := (-\Delta_{\mathbb{T}^d}, F) : F \in \mathcal{F}_{M,L} \}.$$

We write $\mathcal{G}_F$ and $\mathcal{T}_{F,u_0}$ instead of $\mathcal{G}_{\omega_F}$ and $\mathcal{T}_{\omega_F,u_0}$ for simplicity. See Appendix B for the basic facts of this admissible equation class. As in Section 3, let

$$U_{0,R} := \left\{ u_0 \in C(\mathbb{T}^d) : \|u_0\|_{L^\infty(\mathbb{T}^d)} \leqslant R \right\}, \quad U_M := \left\{ u \in C([0,T]; C(\mathbb{T}^d)) : \|u\|_{L^\infty(Q_T)} \leqslant M \right\}.$$

Recall the definition of the Picard map $\mathcal{T}_{F,u_0}$ for $F \in \mathcal{F}_{M,L}$ and $u_0 \in U_{0,R}$:

$$\mathcal{T}_{F,u_0}[v](t) := e^{t\Delta_{\mathbb{T}^d}} u_0 + \int_0^t e^{(t-\tau)\Delta_{\mathbb{T}^d}} F(v(\tau))\, d\tau. \tag{32}$$

For any $R, M > 0$ with $R < M$ and for any $\delta \in (0,1)$, we can take T sufficiently small so that

$$R + TLM \leqslant M, \quad TL \leqslant \delta. \tag{33}$$

Then, by applying Proposition B.1 with $C_S = 1$, we have

$$\|\mathcal{T}_{F,u_0}[v]\|_{L^\infty(Q_T)} \leqslant R + TLM \leqslant M, \quad \|\mathcal{T}_{F,u_0}[v] - \mathcal{T}_{F,u_0}[w]\|_{L^\infty(Q_T)} \leqslant \delta \|v - w\|_{L^\infty(Q_T)}$$

for any $v, w \in U_M$, and

$$\|\mathcal{T}_{F,u_0}[v] - \mathcal{T}_{F,\tilde{u}_0}[v]\|_{L^\infty(Q_T)} \leqslant \|u_0 - \tilde{u}_0\|_{L^\infty(\mathbb{T}^d)}. \tag{34}$$

for any $u_0, \tilde{u}_0 \in U_{0,R}$. Thus, Assumption 3.1 holds with $C_S = 1$ in this setting. Consequently, by Proposition 3.2, for each $F \in \mathcal{F}_{M,L}$, there exists a unique solution operator $\mathcal{G}_F : U_{0,R} \to U_M$ to (31). Define

$$u^{(0)}_{F,u_0} := 0, \quad u^{(\ell+1)}_{F,u_0} := \mathcal{T}_{F,u_0}\big[u^{(\ell)}_{F,u_0}\big], \quad \ell = 0, 1, 2, \cdots. \tag{35}$$

Then, by Lemma 4.1 (Lemma C.1), we have the Picard truncation error bound:

$$\sup_{F\in\mathcal{F}_{M,L}} \sup_{u_0\in U_{0,R}} \left\| u^{(\ell)}_{F,u_0} - \mathcal{G}_F(u_0) \right\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta}. \tag{36}$$

Next, we specify the learning setting. Let $s_0 > d/2$ and

$$U_0 := \left\{u_0 \in H^{s_0}(\mathbb{T}^d) : \|u_0\|_{H^{s_0}(\mathbb{T}^d)} \leqslant R_0\right\} \subset U_{0,R},$$

where R_0 is chosen so that the inclusion holds by the Sobolev embedding. Let Π be a probability measure supported on U_0, and let μ be a probability measure on Q_T. The target equation is generated by some $F_* \in \mathcal{F}_{M,L}$, and we write $\mathcal{G}_* := \mathcal{G}_{F_*}$.

For a candidate operator $\Gamma : U_0 \to U_M$, the ideal population risk is

$$L[\Gamma] := \mathbb{E}_{u_0\sim\Pi,\ z\sim\mu}\left[\left|\Gamma(u_0)(z) - \mathcal{G}_*(u_0)(z)\right|^2\right].$$

The corresponding empirical risk $\widehat{L}_{n,q}$ is the one defined in Section 3.3. Since the Picard iterates considered below are U_M-valued, their pointwise values are uniformly bounded by M.

The abstract ℓ-step Picard class in this torus setting is

$$\mathcal{B}^{\Omega_{\mathrm{eq}}}_\ell = \{b_{F,\ell} : F \in \mathcal{F}_{M,L}\}, \quad b_{F,\ell}(u_0, z) := \left(\mathcal{T}^{[\ell]}_{F,u_0}[0]\right)(z).$$

By (36) and Lemma 4.1 (Lemma C.1), we have

$$\inf_{b\in\mathcal{B}^{\Omega_{\mathrm{eq}}}_\ell} L[b] \leqslant \frac{M^2\delta^{2\ell}}{(1-\delta)^2}.$$

For the Rademacher complexity estimate, we use the equation metric

$$d_{\Omega_{\mathrm{eq}}}(F, G) := \sup_{u_0\in U_{0,R}} \sup_{v\in U_M} \|\mathcal{T}_{F,u_0}[v] - \mathcal{T}_{G,u_0}[v]\|_{L^\infty(Q_T)}.$$

Since the linear part is fixed, the initial-data terms cancel and

$$d_{\Omega_{\mathrm{eq}}}(F, G) \leqslant T\|F - G\|_{L^\infty([-M,M])}.$$

Thus, the entropy of Ω_{eq} is controlled by the entropy of the one-dimensional Lipschitz class $\mathcal{F}_{M,L}$.

Finally, we record the compactness facts used in the implementation theorem. Since $s_0 > d/2$, the Sobolev embedding is compact in the sense that

$$U_0 \Subset C(\mathbb{T}^d).$$

Moreover, by the Ascoli–Arzelà theorem, we have

$$\mathcal{F}_{M,L} \Subset C([-M, M]).$$

These two compactness properties will be used to obtain uniform convergence of the L^∞-stable Fourier cutoffs on the compact families generated by finitely many Picard steps.

F.2 Picard-Type Fourier Neural Operators

We now define the concrete Picard-type FNO implementation used in Theorem 7.2. The construction follows the same Picard-aligned principle as in Furuya et al. [2025], but we use the Fourier diagonalization of the heat semigroup on $\mathbb{T}^d$ and an L^∞-stable Fejér cutoff (i.e. the Cesàro mean of rectangular Fourier partial sums).

Let

$$e_\xi(x) := (2\pi)^{-d/2} e^{i\xi\cdot x}, \quad \xi \in \mathbb{Z}^d.$$

For $f \in C(\mathbb{T}^d)$, we define the Fourier coefficient of f by

$$\widehat{f}(\xi) := \langle \overline{e}_\xi, f\rangle_{\mathbb{T}^d} = \int_{\mathbb{T}^d} f(x)\overline{e_\xi(x)}\, dx.$$

Recall that

$$e^{t\Delta_{\mathbb{T}^d}} f = \sum_{\xi\in\mathbb{Z}^d} e^{-t|\xi|^2} \widehat{f}(\xi) e_\xi,$$

For $N \in \mathbb{N}$, define

$$\mathfrak{m}_N(\xi) := \prod_{r=1}^d \left(1 - \frac{|\xi_r|}{N+1}\right)_+, \quad \xi = (\xi_1, \cdots, \xi_d) \in \mathbb{Z}^d.$$

The symbol $\mathfrak{m}_N$ is fixed and non-trainable. Let

$$\Lambda_N := \left\{\xi \in \mathbb{Z}^d : |\xi_r| \leqslant N,\ r = 1, \cdots, d\right\}.$$

The Fejér operator is

$$P_N f := \sum_{\xi\in\Lambda_N} \mathfrak{m}_N(\xi)\widehat{f}(\xi) e_\xi. \tag{37}$$

Equivalently, P_N is the tensor-product Fejér mean, i.e. the Cesàro mean of rectangular Fourier partial sums. In particular,

$$\|P_N f\|_{L^\infty(\mathbb{T}^d)} \leqslant \|f\|_{L^\infty(\mathbb{T}^d)}, \quad P_N f \to f \quad \text{in } C(\mathbb{T}^d)$$

hold for any $f \in C(\mathbb{T}^d)$ (see, e.g., Grafakos [2014, Chapter 3] and Katznelson [2004, Chapter I]). We define the Fejér approximation of the heat semigroup by

$$S_N(t) f := P_N e^{t\Delta_{\mathbb{T}^d}} f = \sum_{\xi\in\Lambda_N} \mathfrak{m}_N(\xi) e^{-t|\xi|^2} \langle \overline{e}_\xi, f\rangle_{\mathbb{T}^d} e_\xi. \tag{38}$$

For $g \in C([0,T]; C(\mathbb{T}^d))$, define

$$\mathcal{K}_N[g](t) := \int_0^t S_N(t-\tau) g(\tau)\, d\tau.$$

Equivalently, we write

$$\mathcal{K}_N[g](t,x) = \sum_{\xi\in\Lambda_N} \mathfrak{m}_N(\xi) \langle \Psi_{\xi,t}, g\rangle_{Q_T} e_\xi(x),$$

where

$$\Psi_{\xi,t}(\tau, y) := \mathbf{1}_{\{0<\tau<t\}} e^{-(t-\tau)|\xi|^2} \overline{e_\xi(y)}$$

and

$$\langle \Psi_{\xi,t}, g\rangle_{Q_T} := \int_0^T \int_{\mathbb{T}^d} \Psi_{\xi,t}(\tau, y) g(\tau, y)\, dy\, d\tau.$$

Thus, $\mathcal{K}_N$ retains only finitely many spatial Fourier modes, while the time variable remains continuous. Note that the operators P_N, $S_N(t)$, and $\mathcal{K}_N$ preserve real-valued functions, although we use the complex-valued Fourier basis.

Let $H \in \mathbb{N}$. We denote by $\mathcal{N}^{\mathrm{ReLU}}_{H,M,L}$ the class of one-dimensional ReLU networks $\rho : \mathbb{R} \to \mathbb{R}$ with size at most H such that

$$\rho(0) = 0, \quad \mathrm{Lip}_{[-M,M]}(\rho) \leqslant L.$$

Only the network ρ is trainable in the Picard-type FNO below; $\mathfrak{m}_N$ and $e^{-t|\xi|^2}$ are fixed by the PDE structure.

For $\rho \in \mathcal{N}_{H,M,L}^{\mathrm{ReLU}}$ and $u_0 \in U_{0,R}$, define the approximate Picard map by

$$\mathcal{T}_{\rho,u_0}^N[v](t) := S_N(t)u_0 + \mathcal{K}_N[\rho(v)](t), \tag{39}$$

where $\rho(v)(\tau, y) := \rho(v(\tau, y))$. Equivalently, we write

$$\mathcal{T}_{\rho,u_0}^N[v](t,x) = \sum_{\xi\in\Lambda_N} \mathfrak{m}_N(\xi)e^{-t|\xi|^2}\langle \overline{e}_\xi, u_0\rangle_{\mathbb{T}^d} e_\xi(x) + \sum_{\xi\in\Lambda_N} \mathfrak{m}_N(\xi)\langle \Psi_{\xi,t}, \rho(v)\rangle_{Q_T} e_\xi(x).$$

Definition F.1 (Picard-type Fourier neural operator)**.** Let $N, \ell, H \in \mathbb{N}$ and $\rho \in \mathcal{N}_{H,M,L}^{\mathrm{ReLU}}$. For $u_0 \in U_{0,R}$, define

$$u_{N,\rho}^{(0)}[u_0] := 0, \quad u_{N,\rho}^{(j+1)}[u_0] := \mathcal{T}_{\rho,u_0}^N\left[u_{N,\rho}^{(j)}[u_0]\right], \quad j = 0, \cdots, \ell - 1.$$

That is,

$$u_{N,\rho}^{(j+1)}[u_0](t) = S_N(t)u_0 + \mathcal{K}_N\left[\rho\left(u_{N,\rho}^{(j)}[u_0]\right)\right](t).$$

The ℓ-step Picard-type Fourier neural operator $\Gamma_{N,\ell,\rho}^{\mathrm{FNO}} : U_0 \to U_M$ is defined by

$$\Gamma_{N,\ell,\rho}^{\mathrm{FNO}}(u_0) := u_{N,\rho}^{(\ell)}[u_0]. \tag{40}$$

The corresponding implementation class is

$$\mathfrak{N}_{N,\ell,H}^{\mathrm{FNO}}(M, L) := \{\Gamma_{N,\ell,\rho}^{\mathrm{FNO}} : \rho \in \mathcal{N}_{H,M,L}^{\mathrm{ReLU}}\}$$

as a class of maps from U_0 to U_M.

Following the two-channel construction in the proof of Lemma 5 of Furuya et al. [2025], we express the operator in Definition F.1 as a composition of an input map, shared hidden blocks, and an output map. Define the input and output maps by

$$\mathcal{L}_N u_0 := \begin{pmatrix} S_N(\cdot)u_0 \\ S_N(\cdot)u_0 \end{pmatrix}, \quad \mathcal{Q}(v_1, v_2) := v_1.$$

For $v = (v_1, v_2)^\top$, set

$$\mathcal{R}_\rho(v) := \begin{pmatrix} \rho(v_1) \\ v_2 \end{pmatrix}, \quad W := \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbb{K}_N v := \begin{pmatrix} \mathcal{K}_N[v_1] \\ 0 \end{pmatrix}.$$

Then the hidden block $\mathcal{B}_{N,\rho} := (W + \mathbb{K}_N) \circ \mathcal{R}_\rho$ satisfies

$$\mathcal{B}_{N,\rho}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_2 + \mathcal{K}_N[\rho(v_1)] \\ v_2 \end{pmatrix}.$$

Since $\rho(0) = 0$, the first Picard iterate is $u_{N,\rho}^{(1)}[u_0] = S_N(\cdot)u_0$. Each hidden block performs one additional Picard step while preserving $S_N(\cdot)u_0$ in the second channel. Hence

$$\Gamma_{N,\ell,\rho}^{\mathrm{FNO}} = \mathcal{Q} \circ \mathcal{B}_{N,\rho}^{[\ell-1]} \circ \mathcal{L}_N,$$

where $\mathcal{B}_{N,\rho}^{[0]}$ is the identity map. The same scalar network ρ is shared across all hidden blocks, while the linear operators are fixed.

Remark F.2 (Comparison with other neural-operator architectures). A standard FNO layer has the form

$$v_{j+1} = \sigma\left(W_j v_j + \mathcal{F}_x^{-1}[R_j \mathcal{F}_x v_j] + b_j\right),$$

where W_j, R_j, and b_j are trainable and may vary between layers; see Li et al. [2021]. The construction in the proof of Lemma 5 of Furuya et al. [2025] instead follows Picard iteration. It uses two channels to retain the current iterate and the linear initial-data term, and shares the same nonlinear network across Picard steps. The integral operator is approximated by a finite-rank space-time kernel expansion. Our construction uses the same two-channel structure. The first channel is updated by

$$v_{j+1} = a + \mathcal{K}_N[\rho(v_j)], \quad a := S_N(\cdot)u_0,$$

while the second channel retains a. Only the scalar network ρ is trainable. The input map, output map, and Fourier–Duhamel operator are fixed, and the same ρ is shared across all blocks. Here, the Fourier cutoff acts only in space, while the Volterra integral in time is retained. Thus, each block performs one Picard step rather than a standard FNO layer.

Remark F.3. Assume that the parameters R, M, L, T, δ satisfy (33). Then the following elementary stability property holds, which shows that no additional clipping is required in Definition F.1. In fact, since both $e^{t\Delta_{\mathbb{T}^d}}$ and P_N are L^∞-contractive, we have

$$\|S_N(t)f\|_{L^\infty(\mathbb{T}^d)} \leqslant \|f\|_{L^\infty(\mathbb{T}^d)}, \quad \|\mathcal{K}_N[g]\|_{L^\infty(Q_T)} \leqslant T\|g\|_{L^\infty(Q_T)}. \tag{41}$$

Hence, for any $u_0 \in U_{0,R}$ and $v \in U_M$,

$$\|\mathcal{T}^N_{\rho,u_0}[v]\|_{L^\infty(Q_T)} \leqslant R + TLM \leqslant M. \tag{42}$$

Moreover, for $v, w \in U_M$,

$$\begin{aligned} \|\mathcal{T}^N_{\rho,u_0}[v] - \mathcal{T}^N_{\rho,u_0}[w]\|_{L^\infty(Q_T)} &\leqslant T\,\mathrm{Lip}_{[-M,M]}(\rho)\|v-w\|_{L^\infty(Q_T)} \\ &\leqslant \delta\|v-w\|_{L^\infty(Q_T)}. \end{aligned} \tag{43}$$

Therefore $\mathcal{T}^N_{\rho,u_0} : U_M \to U_M$ is a δ-contraction, uniformly in N, $\rho \in \mathcal{N}^{\mathrm{ReLU}}_{H,M,L}$, and $u_0 \in U_{0,R}$. In particular, $\Gamma^{\mathrm{FNO}}_{N,\ell,\rho}(u_0) \in U_M$ for any $u_0 \in U_{0,R}$.

Remark F.4 (Trainable parameters and computational interpretation). We clarify the implementation meaning of Definition F.1.

(a) The trainable part of the Picard-type FNO is only the scalar ReLU network $\rho \in \mathcal{N}^{\mathrm{ReLU}}_{H,M,L}$. The parameters N, ℓ, H are architecture or approximation parameters, not trainable weights. The Fourier multiplier mask $\mathfrak{m}_N(\xi)$, the Fourier basis e_ξ, and the heat factors $e^{-t|\xi|^2}$ are fixed by the PDE structure.

(b) The dual-pair terms are not additional trainable parameters. For example, $\langle \overline{e}_\xi, u_0\rangle_{\mathbb{T}^d}$ is the Fourier coefficient of the input, and $\langle \Psi_{\xi,t}, g\rangle_{Q_T}$ is the coefficient of the fixed Fourier-heat kernel applied to g. These quantities are computed in the forward pass. They depend on the current iterate through $g = \rho(u^{(j)}_{N,\rho}[u_0])$, and therefore gradients flow through them to the parameters of ρ, but the kernels themselves are not learned.

(c) On a periodic spatial grid, the spatial dual pairs can be computed by FFT. The Duhamel term is then evaluated mode by mode as an exponential convolution in time,

$$\int_0^t e^{-(t-\tau)|\xi|^2}\widehat{g}(\tau,\xi)\,d\tau,$$

which can be approximated by quadrature or by a time-stepping recurrence on an equally spaced time grid. Thus, the dual-pair notation is an analytic way of writing a computable Fourier-heat convolution.

(d) The numerical cost is not dimension-free. If m^d is the number of spatial grid points, n_t is the number of time points, and ℓ is the Picard depth, then a standard grid implementation has a leading spatial FFT cost of order

$$O\big(\ell\, n_t\, m^d \log(m^d)\big)\,,$$

up to the additional cost of the time convolution. Equivalently, if N denotes the cutoff level in each coordinate direction, then the tensor-product Fourier index set satisfies $|\Lambda_N| = (2N+1)^d \asymp N^d$. Thus, the usual numerical cost of high-dimensional spectral discretization remains.

(e) The role of the theory is different from eliminating all numerical dimension-dependence. The Fourier cutoff N enters the error bound only through the implementation error, while the Rademacher complexity is controlled by the abstract Picard class and, in the torus example, by the scalar nonlinearity class $\mathcal{F}_{M,L}$. Hence, increasing the Fourier rank improves the implementation error without enlarging the statistical complexity term.

(f) The constraints $\rho(0) = 0$ and $\mathrm{Lip}_{[-M,M]}(\rho) \leqslant L$ are part of the theoretical architecture class. They ensure that the approximate Picard map remains a self-map on U_M and is a δ-contraction under (33). In practical training, one should enforce or approximate these constraints, for instance by using a slope-constrained piecewise affine parameterization, projection, or normalization. Without these constraints, the no-clipping stability argument need not apply.

F.3 Proof of Theorem 7.2

In this appendix, we prove the implementation error estimate for the Picard-type FNO. The argument follows the Picard-iteration strategy of Furuya et al. [2025], but the rank-N part is handled differently. Instead of postulating an explicit rank-N approximation rate for the Green kernel, we use the compactness of U_0 in $C(\mathbb{T}^d)$ and the compactness of $\mathcal{F}_{M,L}$ in $C([-M,M])$ to obtain uniform convergence of the L^∞-stable Fejér cutoffs on the compact families generated by finitely many Picard steps.

Theorem F.5 (Theorem 7.2, Implementation error for the Picard-type FNO)**.** *Let $s_0 > d/2$ and let $U_0 = \{u_0 \in H^{s_0}(\mathbb{T}^d) : \|u_0\|_{H^{s_0}} \leqslant R_0\} \subset U_{0,R}$. Assume* (33)*. Then, for any $\eta \in (0,1)$ and every $F \in \mathcal{F}_{M,L}$, there exists a one-dimensional ReLU network ρ_F such that*

$$\|F - \rho_F\|_{L^\infty([-M,M])} \leqslant \eta, \quad \rho_F(0) = 0, \quad \mathrm{Lip}_{[-M,M]}(\rho_F) \leqslant L, \tag{44}$$

and

$$H(\rho_F) \leqslant C(1 + ML\eta^{-1}), \tag{45}$$

where $C > 0$ is a universal constant. Moreover, for each fixed $\ell \in \mathbb{N}$, there exists a sequence $a_N(\ell) \to 0$ as $N \to \infty$ such that

$$\sup_{F \in \mathcal{F}_{M,L}} \sup_{u_0 \in U_0} \left\| \Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}(u_0) - \mathcal{T}^{[\ell]}_{F,u_0}[0] \right\|_{L^\infty(Q_T)} \leqslant \frac{1-\delta^\ell}{1-\delta}\, T\eta + a_N(\ell). \tag{46}$$

Here, the Fourier truncation error $a_N(\ell)$ can be chosen independently of η. Consequently, if

$$\mathrm{embed}_{N,\eta}\big(b_{F,\ell}\big) := \Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}, \quad b_{F,\ell}(u_0, z) := \left(\mathcal{T}^{[\ell]}_{F,u_0}[0]\right)(z),$$

then the implementation error

$$\varepsilon^{\mathrm{FNO}}_{\mathrm{imp}}(N,\ell,\eta) := \sup_{b \in \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} |L[\mathrm{embed}_{N,\eta}(b)] - L[b]|$$

satisfies

$$\varepsilon^{\mathrm{FNO}}_{\mathrm{imp}}(N,\ell,\eta) \leqslant \sup_{F \in \mathcal{F}_{M,L}} \left|L\big[\Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}\big] - L[b_{F,\ell}]\right| \leqslant 4M \left\{ \frac{1-\delta^\ell}{1-\delta}\, T\eta + a_N(\ell) \right\}. \tag{47}$$

Proof. The proof of (46) is divided into three steps.

Step 1: Approximation of the nonlinearity. Fix $F \in \mathcal{F}_{M,L}$. If $L = 0$, then, since $F(0) = 0$, we have $F \equiv 0$ on $[-M, M]$, and we simply take $\rho_F \equiv 0$.

Suppose $L > 0$. Choose a partition

$$-M = x_0 < x_1 < \cdots < x_K = M$$

which contains 0 and satisfies

$$h := \max_{0 \leqslant k \leqslant K-1} |x_{k+1} - x_k| \leqslant \frac{\eta}{L}, \quad K \leqslant C(1 + ML\eta^{-1})$$

with a universal constant C. Let ρ_F be the continuous piecewise affine interpolation of F on this partition. The ρ_F satisfies (44). In fact, we have $\rho_F(0) = 0$ since the partition contains 0 and $F(0) = 0$, and $\mathrm{Lip}_{[-M,M]}(\rho_F) \leqslant L$ since every secant slope of F on the partition has absolute value at most L. Moreover, for $x \in [x_k, x_{k+1}]$, writing $x = (1-\theta)x_k + \theta x_{k+1}$ with $0 \leqslant \theta \leqslant 1$, we have

$$|F(x) - \rho_F(x)| \leqslant (1-\theta)|F(x) - F(x_k)| + \theta|F(x) - F(x_{k+1})| \leqslant Lh.$$

Therefore,

$$\|F - \rho_F\|_{L^\infty([-M,M])} \leqslant Lh \leqslant \eta.$$

Thus, (44) is shown.

By the ReLU network representation of continuous piecewise linear functions, it is seen that ρ_F is represented by a ReLU network with (45) (see the proof of Proposition 1 in Yarotsky [2017]).

Since $\rho_F(0) = 0$ and $\mathrm{Lip}_{[-M,M]}(\rho_F) \leqslant L$, the estimates used for the Picard map remain valid with F replaced by ρ_F. Consequently, the approximate Picard map $\mathcal{T}^N_{\rho_F,u_0}$ defined in (39) satisfies (42) and (43) for all $v, w \in U_M$.

To prove (46), we decompose

$$\begin{aligned}\left\|\Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}(u_0) - \mathcal{T}^{[\ell]}_{F,u_0}[0]\right\|_{L^\infty(Q_T)} &= \left\|u^{(\ell)}_{N,\rho}[u_0] - u^{(\ell)}_{F,u_0}\right\|_{L^\infty(Q_T)} \\ &\leqslant \left\|u^{(\ell)}_{N,\rho}[u_0] - \widetilde{u}^{(\ell)}_{\rho_F,u_0}\right\|_{L^\infty(Q_T)} + \left\|\widetilde{u}^{(\ell)}_{\rho_F,u_0} - u^{(\ell)}_{F,u_0}\right\|_{L^\infty(Q_T)},\end{aligned} \tag{48}$$

where $u^{(\ell)}_{F,u_0}$ and $u^{(\ell)}_{N,\rho}[u_0]$ are defined by (35) and (40), respectively. Here, $\widetilde{u}^{(\ell)}_{\rho_F,u_0}$ is defined by

$$\widetilde{u}^{(0)}_{\rho_F,u_0} := 0, \quad \widetilde{u}^{(j+1)}_{\rho_F,u_0} := \mathcal{T}_{\rho_F,u_0}\big[\widetilde{u}^{(j)}_{\rho_F,u_0}\big],$$

where

$$\mathcal{T}_{\rho_F,u_0}[v](t) := e^{t\Delta_{\mathbb{T}^d}}u_0 + \int_0^t e^{(t-\tau)\Delta_{\mathbb{T}^d}}\rho_F(v(\tau))\,d\tau.$$

Step 2: Error caused by replacing F with ρ_F. In this step, we prove

$$\sup_{F\in\mathcal{F}_{M,L}}\sup_{u_0\in U_0}\left\|\widetilde{u}^{(\ell)}_{\rho_F,u_0} - u^{(\ell)}_{F,u_0}\right\|_{L^\infty(Q_T)} \leqslant \frac{1-\delta^\ell}{1-\delta}\,T\eta. \tag{49}$$

Since $e^{t\Delta_{\mathbb{T}^d}}$ is L^∞-contractive and (44) holds, we estimate

$$\begin{aligned}&\left\|\widetilde{u}^{(j+1)}_{\rho_F,u_0} - u^{(j+1)}_{F,u_0}\right\|_{L^\infty(Q_T)} \\ &\quad\leqslant T\left\|\rho_F\left(\widetilde{u}^{(j)}_{\rho_F,u_0}\right) - F\left(u^{(j)}_{F,u_0}\right)\right\|_{L^\infty(Q_T)} \\ &\quad\leqslant T\left\|\rho_F\left(\widetilde{u}^{(j)}_{\rho_F,u_0}\right) - \rho_F\left(u^{(j)}_{F,u_0}\right)\right\|_{L^\infty(Q_T)} + T\left\|\rho_F\left(u^{(j)}_{F,u_0}\right) - F\left(u^{(j)}_{F,u_0}\right)\right\|_{L^\infty(Q_T)} \\ &\quad\leqslant \delta\left\|\widetilde{u}^{(j)}_{\rho_F,u_0} - u^{(j)}_{F,u_0}\right\|_{L^\infty(Q_T)} + T\eta.\end{aligned}$$

Since the zeroth iterates coincide, the induction argument yields (49).

Step 3: Uniform Fourier truncation error by compactness. In this step, we prove that for each fixed $\ell \in \mathbb{N}$, there exists a sequence $a_N(\ell) \to 0$ as $N \to \infty$ such that

$$\sup_{F\in\mathcal{F}_{M,L}}\sup_{u_0\in U_0}\left\|u^{(\ell)}_{N,\rho_F}[u_0] - \widetilde{u}^{(\ell)}_{\rho_F,u_0}\right\|_{L^\infty(Q_T)} \leqslant a_N(\ell). \tag{50}$$

For $G \in \mathcal{F}_{M,L}$, we define $\{u^{(j)}_{G,u_0}\}_j$ by (35), and the Fejér-truncated Picard iterates by

$$u^{(0)}_{N,G,u_0} := 0, \quad u^{(j+1)}_{N,G,u_0} := S_N(\cdot)u_0 + \mathcal{K}_N\left[G\left(u^{(j)}_{N,G,u_0}\right)\right] \tag{51}$$

(note that $u^{(\ell)}_{N,G}[u_0] = u^{(\ell)}_{N,G,u_0}$). Since $G(0) = 0$ and $\mathrm{Lip}_{[-M,M]}(G) \leqslant L$, the same estimate as (42) and (43) shows that the truncated Picard map is a δ-contraction from U_M into itself. Thus, all iterates in (51) are U_M-valued.

First, we prove that the sets

$$\mathcal{C}_j := \left\{u^{(j)}_{G,u_0} : G \in \mathcal{F}_{M,L},\ u_0 \in U_0\right\} \subset C([0,T]; C(\mathbb{T}^d))$$

are compact for $j = 0, 1, \ldots, \ell$. To this end, we show by induction on j that the map

$$(u_0, G) \mapsto u^{(j)}_{G,u_0}$$

is continuous. For $j = 0$, this map is constant. Suppose that it is continuous for some fixed j. The map

$$(u_0, G, u) \mapsto e^{t\Delta_{\mathbb{T}^d}}u_0 + \int_0^t e^{(t-\tau)\Delta_{\mathbb{T}^d}}G(u(\tau))\,d\tau$$

is continuous from $U_0 \times \mathcal{F}_{M,L} \times U_M$ into $C([0,T]; C(\mathbb{T}^d))$. Indeed, $u_0 \mapsto e^{(\cdot)\Delta_{\mathbb{T}^d}} u_0$ is continuous by the L^∞-contractivity of the heat semigroup, and

$$\|G(u) - \widetilde{G}(\widetilde{u})\|_{L^\infty(Q_T)} \leqslant \|G - \widetilde{G}\|_{L^\infty([-M,M])} + L\|u - \widetilde{u}\|_{L^\infty(Q_T)}. \tag{52}$$

Therefore, the Picard recurrence and the induction hypothesis show that $(u_0, G) \mapsto u^{(j+1)}_{G,u_0}$ is continuous. This proves continuity by induction. Since U_0 and $\mathcal{F}_{M,L}$ are compact in $C(\mathbb{T}^d)$ and $C([-M,M])$, respectively, each $\mathcal{C}_j$ is compact as the continuous image of $U_0 \times \mathcal{F}_{M,L}$.

Next, we define

$$\mathcal{D}_j := \{G(u) : G \in \mathcal{F}_{M,L},\ u \in \mathcal{C}_j\} \subset C([0,T]; C(\mathbb{T}^d)).$$

Then $\mathcal{D}_j$ is compact for any fixed j by the same continuity estimate (52).

Since P_N is the Fejér operator (37), the convergence $P_N f \to f$ in $C(\mathbb{T}^d)$ holds for each $f \in C(\mathbb{T}^d)$. Hence, by the standard compactness argument, we have

$$\sup_{f \in \mathcal{K}} \|P_N f - f\|_{L^\infty(\mathbb{T}^d)} \to 0 \quad \text{for any compact } K \subset C(\mathbb{T}^d). \tag{53}$$

Applying (53) with $K = \mathcal{H}_0$ given by

$$\mathcal{H}_0 := \left\{ e^{t\Delta_{\mathbb{T}^d}} u_0 : u_0 \in U_0,\ t \in [0,T] \right\},$$

where $\mathcal{H}_0$ is compact because U_0 is compact in $C(\mathbb{T}^d)$ and the map $(u_0, t) \mapsto e^{t\Delta_{\mathbb{T}^d}} u_0$ is continuous from $U_0 \times [0,T]$ into $C(\mathbb{T}^d)$, we obtain

$$\alpha_N := \sup_{u_0 \in U_0} \left\| S_N(\cdot) u_0 - e^{(\cdot)\Delta_{\mathbb{T}^d}} u_0 \right\|_{L^\infty(Q_T)} \to 0$$

as $N \to \infty$. For each j, apply (53) to the compact set

$$\mathcal{H}_j := \left\{ e^{(t-\tau)\Delta_{\mathbb{T}^d}} g(\tau) : g \in \mathcal{D}_j,\ 0 \leqslant \tau \leqslant t \leqslant T \right\} \subset C(\mathbb{T}^d).$$

This yields

$$\beta_{N,j} := \sup_{g \in \mathcal{D}_j} \|\mathcal{K}_N[g] - \mathcal{K}[g]\|_{L^\infty(Q_T)} \leqslant T \sup_{h \in \mathcal{H}_j} \|P_N h - h\|_{L^\infty(\mathbb{T}^d)} \to 0$$

as $N \to \infty$ for each j, where

$$\mathcal{K}[g](t) := \int_0^t e^{(t-\tau)\Delta_{\mathbb{T}^d}} g(\tau)\, d\tau.$$

Let

$$d_{N,j} := \sup_{G \in \mathcal{F}_{M,L}} \sup_{u_0 \in U_0} \left\| u^{(j)}_{N,G,u_0} - u^{(j)}_{G,u_0} \right\|_{L^\infty(Q_T)}.$$

Then $d_{N,0} = 0$. Using (33), (35), (41), (51), and (52), we obtain

$$d_{N,j+1} \leqslant \alpha_N + \beta_{N,j} + TL\, d_{N,j} \leqslant \alpha_N + \beta_{N,j} + \delta d_{N,j}. \tag{54}$$

Indeed, the term $\beta_{N,j}$ controls $(\mathcal{K}_N - \mathcal{K})[G(u^{(j)}_{G,u_0})]$, and the last term controls $\mathcal{K}_N[G(u^{(j)}_{N,G,u_0}) - G(u^{(j)}_{G,u_0})]$. Iterating (54) gives

$$d_{N,\ell} \leqslant \sum_{j=0}^{\ell-1} \delta^{\ell-1-j} \left(\alpha_N + \beta_{N,j}\right) =: a_N(\ell). \tag{55}$$

Since ℓ is fixed, $\alpha_N \to 0$ and $\beta_{N,j} \to 0$ for each $j = 0, \ldots, \ell - 1$, we have

$$a_N(\ell) \to 0 \quad \text{as } N \to \infty.$$

Taking $G = \rho_F$, and recalling that $\rho_F \in \mathcal{F}_{M,L}$ on $[-M,M]$, we obtain (50), where $u^{(\ell)}_{N,\rho_F,u_0} = \Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}(u_0)$ by Definition F.1.

Combining (48), (49) and (50), we conclude

$$\sup_{F\in\mathcal{F}_{M,L}}\sup_{u_0\in U_0}\left\|\Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}(u_0)-\mathcal{T}^{[\ell]}_{F,u_0}[0]\right\|_{L^\infty(Q_T)}\leqslant\frac{1-\delta^\ell}{1-\delta}\,T\eta+a_N(\ell).$$

This proves (46).

Finally, let $b_{F,\ell}(u_0,z)=(\mathcal{T}^{[\ell]}_{F,u_0}[0])(z)$ and $\Gamma_F:=\Gamma^{\mathrm{FNO}}_{N,\ell,\rho_F}$. Since both $b_{F,\ell}$ and Γ_F take values in $[-M,M]$, and since $\mathcal{G}_*(u_0)(z)\in[-M,M]$, we have

$$\begin{aligned}&\left||\Gamma_F(u_0)(z)-\mathcal{G}_*(u_0)(z)|^2-|b_{F,\ell}(u_0,z)-\mathcal{G}_*(u_0)(z)|^2\right|\\&\qquad\leqslant 4M\left|\Gamma_F(u_0)(z)-b_{F,\ell}(u_0,z)\right|.\end{aligned}$$

Taking expectation and then the supremum over $F\in\mathcal{F}_{M,L}$ gives (47). The proof is complete. □

Proposition F.6 (Explicit Fourier truncation rate)**.** *Assume the setting of Theorem F.5. For any $0<\gamma<\min\{s_0-d/2,1\}$, there exists a constant $C_{\gamma,d,s_0}>0$ such that*

$$a_N(\ell)\leqslant C_{\gamma,d,s_0}\frac{1-\delta^\ell}{1-\delta}\left(R_0+LMT^{1-\frac{\gamma}{2}}\right)N^{-\gamma}\tag{56}$$

for any $N,\ell\in\mathbb{N}$. In particular, this bound is independent of η and uniform in the Picard depth ℓ.

Proof. Write $S(t):=e^{t\Delta_{\mathbb{T}^d}}$. For $0<\gamma<1$, the tensor-product Fejér operator satisfies

$$\|(P_N-I)f\|_{L^\infty}\leqslant C_{\gamma,d}N^{-\gamma}[f]_{C^{0,\gamma}},\tag{57}$$

where $[f]_{C^{0,\gamma}}$ denotes the Hölder seminorm on $\mathbb{T}^d$. Indeed, the normalized one-dimensional Fejér kernel K_N satisfies

$$K_N(y)\leqslant C\min\{N+1,(N+1)^{-1}|y|^{-2}\},\quad 0<|y|\leqslant\pi.$$

Hence

$$\int_{-\pi}^{\pi}|y|^\gamma K_N(y)\,dy\leqslant C_\gamma N^{-\gamma}.$$

Applying this estimate to the product kernel gives (57).

Since $\gamma<s_0-d/2$, the Sobolev embedding gives

$$[u_0]_{C^{0,\gamma}}\leqslant C_{\gamma,d,s_0}\|u_0\|_{H^{s_0}}.$$

The heat semigroup preserves the Hölder seminorm. Thus, the initial-data error defined above satisfies

$$\alpha_N\leqslant C_{\gamma,d,s_0}R_0N^{-\gamma}.$$

Next, the heat semigroup satisfies

$$[S(s)f]_{C^{0,\gamma}}\leqslant C_{\gamma,d}s^{-\gamma/2}\|f\|_{L^\infty},\quad s>0.$$

This follows from L^∞-contractivity and the gradient estimate $\|\nabla S(s)f\|_{L^\infty}\leqslant C_ds^{-1/2}\|f\|_{L^\infty}$. Combining this with (57), we obtain

$$\|(P_N-I)S(s)f\|_{L^\infty}\leqslant C_{\gamma,d}N^{-\gamma}s^{-\gamma/2}\|f\|_{L^\infty}.$$

For any $g\in\mathcal{D}_j$, we have $\|g\|_{L^\infty(Q_T)}\leqslant LM$. Therefore,

$$\begin{aligned}\|(\mathcal{K}_N-\mathcal{K})[g]\|_{L^\infty(Q_T)}&\leqslant C_{\gamma,d}LMN^{-\gamma}\sup_{0<t\leqslant T}\int_0^t(t-\tau)^{-\gamma/2}\,d\tau\\&\leqslant C_{\gamma,d}LMT^{1-\gamma/2}N^{-\gamma}.\end{aligned}$$

Hence

$$\beta_{N,j}\leqslant C_{\gamma,d}LMT^{1-\gamma/2}N^{-\gamma}$$

uniformly in j. Substituting these estimates into (55) proves (56). □

Remark F.7 (Compactness and quantitative convergence). (a) The above proof differs from the Green-kernel approximation argument of Furuya et al. [2025], where the approximation error is controlled through a chosen kernel expansion. Here, we use the compactness of U_0 and $\mathcal{F}_{M,L}$ and the uniform convergence of Fejér cutoffs on the resulting compact families to obtain $a_N(\ell) \to 0$ for any fixed ℓ.

(b) Proposition F.6 gives an explicit rate under the same assumptions, using the Sobolev regularity of the initial data and heat-semigroup smoothing. Thus, the implementation error can be made arbitrarily small by choosing $\eta > 0$ sufficiently small and N sufficiently large, uniformly in ℓ. Increasing ℓ also reduces the Picard truncation error.

(c) The Fourier cutoff N enters only through the implementation error and does not affect the Rademacher complexity of the abstract Picard class. In this sense, the bound avoids the curse of parametric complexity at the statistical level, although the number of Fourier modes and the computational cost still depend on the spatial dimension.

F.4 Proof of Lemma 7.3

We next estimate the Rademacher complexity of the abstract Picard class in the torus setting. The key point is that the entropy is measured at the level of the scalar nonlinearity class $\mathcal{F}_{M,L}$, and not at the level of the Fourier-rank implementation.

Lemma F.8 (Lemma 7.3, Rademacher complexity of the Picard class)**.** *Let $\mathcal{S}_{n,q}^{U_0} = \{(u_{0,i}, \{z_{i,j}\}_{j=1}^q)\}_{i=1}^n$ be a trajectory-level sample as in Section 3.3. Assume* (33)*. Then, for any $\ell \in \mathbb{N}$, the trajectory-normalized empirical Rademacher complexity of the Picard class satisfies*

$$\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} = \widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{U_0}}\big(\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}\big) \leqslant CM\sqrt{\frac{LT(1-\delta^\ell)}{(1-\delta)n}}, \tag{58}$$

where $C > 0$ is a universal constant. In particular,

$$\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} \leqslant CM\sqrt{\frac{LT}{(1-\delta)n}}, \tag{59}$$

so the Rademacher term is uniformly bounded in the Picard depth ℓ and is independent of the Fourier cutoff N of the Picard-type FNO implementation.

Proof. The case $\ell = 0$ is trivial, since $\mathcal{B}_0^{\Omega_{\mathrm{eq}}}$ contains only the zero predictor. We therefore assume $\ell \geqslant 1$. Let

$$d_{\mathcal{S},2}(b,b') := \left(\frac{1}{nq}\sum_{i=1}^n\sum_{j=1}^q \big|b(u_{0,i},z_{i,j}) - b'(u_{0,i},z_{i,j})\big|^2\right)^{1/2}$$

be the empirical pointwise metric on predictors. By the trajectory-normalized Dudley entropy bound used in Section 5,

$$\widehat{\mathfrak{R}}_{\mathcal{S}_{n,q}^{U_0}}\big(\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}\big) \leqslant \frac{C}{\sqrt{n}}\int_0^{2M}\sqrt{\log N\left(\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}, d_{\mathcal{S},2}, \varepsilon\right)}\,d\varepsilon. \tag{60}$$

Here, every element of $\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}$ is $[-M,M]$-valued, since the Picard iterates stay in U_M. The factor $n^{-1/2}$ comes from the trajectory-normalized complexity defined in Section 5.

We first reduce the covering number of the Picard class to that of the scalar nonlinearity class. Let $F, G \in \mathcal{F}_{M,L}$ and fix $u_0 \in U_{0,R}$. Define

$$u_F^{(0)} = u_G^{(0)} := 0, \quad u_F^{(j+1)} := \mathcal{T}_{F,u_0}\big[u_F^{(j)}\big], \quad u_G^{(j+1)} := \mathcal{T}_{G,u_0}\big[u_G^{(j)}\big].$$

Both sequences are U_M-valued. Since the heat semigroup is L^∞-contractive and F is L-Lipschitz on $[-M,M]$, we have

$$\begin{aligned}\|u_F^{(j+1)} - u_G^{(j+1)}\|_{L^\infty(Q_T)} &\leqslant T\left\|F(u_F^{(j)}) - G(u_G^{(j)})\right\|_{L^\infty(Q_T)}\\ &\leqslant \delta\|u_F^{(j)} - u_G^{(j)}\|_{L^\infty(Q_T)} + T\|F-G\|_{L^\infty([-M,M])}.\end{aligned}$$

Since the zeroth iterates coincide, induction gives

$$\left\|\mathcal{T}_{F,u_0}^{[\ell]}[0]-\mathcal{T}_{G,u_0}^{[\ell]}[0]\right\|_{L^\infty(Q_T)} \leqslant T\frac{1-\delta^\ell}{1-\delta}\|F-G\|_{L^\infty([-M,M])}.$$

Therefore, for any sample $\mathcal{S}_{n,q}^{U_0}$,

$$N\left(\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}, d_{\mathcal{S},2}, \varepsilon\right) \leqslant N\left(\mathcal{F}_{M,L}, \|\cdot\|_{L^\infty([-M,M])}, \frac{(1-\delta)\varepsilon}{T(1-\delta^\ell)}\right). \tag{61}$$

We next use the elementary entropy estimate for the one-dimensional Lipschitz class:

$$\log N\left(\mathcal{F}_{M,L}, \|\cdot\|_{L^\infty([-M,M])}, r\right) \leqslant C\frac{ML}{r}, \quad r>0. \tag{62}$$

For completeness, we recall the short proof. If $L=0$, then $\mathcal{F}_{M,L}=\{0\}$. Assume $L>0$. Choose a grid on $[-M,M]$ containing 0 with mesh size $h \asymp r/L$. The number of grid intervals is $K \leqslant CML/r$. Since $F(0)=0$ and $\mathrm{Lip}_{[-M,M]}(F) \leqslant L$, the values at adjacent grid points differ by at most $Lh \asymp r$. Quantizing the grid values with mesh comparable to r and using the adjacent-increment constraint yields at most C^K admissible quantized sequences. The corresponding piecewise affine interpolants form an r-net in $L^\infty([-M,M])$. Hence $\log N \leqslant CK \leqslant CML/r$.

Combining (61) and (62), we obtain

$$\log N\left(\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}, d_{\mathcal{S},2}, \varepsilon\right) \leqslant C\frac{MLT(1-\delta^\ell)}{(1-\delta)\varepsilon}.$$

Substituting this into (60) gives

$$\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} \leqslant \frac{C}{\sqrt{n}}\int_0^{2M}\sqrt{\frac{MLT(1-\delta^\ell)}{(1-\delta)\varepsilon}}\,d\varepsilon \leqslant CM\sqrt{\frac{LT(1-\delta^\ell)}{(1-\delta)n}}.$$

This proves (58), and hence, (59) follows since $1-\delta^\ell \leqslant 1$. The proof of Lemma F.8 is finished. □

F.5 Proof of Theorem 7.4

We next combine the implementation estimate for the Picard-type FNO with the implementation-agnostic generalization error bound of Theorem 5.1.

Theorem F.9 (Theorem 7.4, Generalization error bound for the Picard-type FNO)**.** *Assume the setting of Appendix F.1 and Theorem F.5. Let $\eta \in (0,1)$ and $\ell, N \in \mathbb{N}$. For each $F \in \mathcal{F}_{M,L}$, choose a ReLU network ρ_F as in Theorem F.5, and define*

$$\mathrm{embed}_{N,\eta}\left(b_{F,\ell}\right) := \Gamma_{N,\ell,\rho_F}^{\mathrm{FNO}}. \tag{63}$$

Let

$$\widehat{b}_\ell \in \operatorname*{argmin}_{b\in\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} \widehat{L}_{n,q}[b],$$

and set

$$\widehat{\Gamma}_{N,\ell,\eta}^{\mathrm{FNO}} := \mathrm{embed}_{N,\eta}(\widehat{b}_\ell).$$

Then there exists a universal constant $C>0$ such that, for any $0<\zeta<1$, with probability at least $1-\zeta$ over the trajectory-level sample $\mathcal{S}_{n,q}^{U_0}$,

$$\begin{aligned} L\left[\widehat{\Gamma}_{N,\ell,\eta}^{\mathrm{FNO}}\right] \leqslant 4M\left\{\frac{1-\delta^\ell}{1-\delta}\,T\eta + a_N(\ell)\right\} + \frac{M^2\delta^{2\ell}}{(1-\delta)^2} \\ + CM^2\sqrt{\frac{LT(1-\delta^\ell)}{(1-\delta)n}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}. \end{aligned} \tag{64}$$

Here, $a_N(\ell)\to 0$ as $N\to\infty$ is the Fourier implementation error from Theorem 7.2. In particular, the statistical term is uniformly bounded in the Picard depth ℓ and is independent of the Fourier cutoff N.

Proof. The proof is a direct specialization of Theorem 5.1 to the torus heat equation class.

First, by Theorem F.5 (Theorem 7.2), the map $\mathrm{embed}_{N,\eta}$ defined in (63) implements the abstract Picard class $\mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}$ by Picard-type FNOs. More precisely, for every $F \in \mathcal{F}_{M,L}$,

$$\sup_{u_0 \in U_0} \|\mathrm{embed}_{N,\eta}(b_{F,\ell})(u_0) - b_{F,\ell}(u_0,\cdot)\|_{L^\infty(Q_T)} \leqslant \frac{1-\delta^\ell}{1-\delta}\, T\eta + a_N(\ell).$$

Since both $b_{F,\ell}$ and $\mathrm{embed}_{N,\eta}(b_{F,\ell})$ are $[-M,M]$-valued, and since $\mathcal{G}_*(u_0)(z) \in [-M,M]$, the squared loss is $4M$-Lipschitz in the prediction variable on this range. Therefore,

$$\varepsilon_{\mathrm{imp}}^{\mathrm{FNO}}(N,\eta) := \sup_{b \in \mathcal{B}_\ell^{\Omega_{\mathrm{eq}}}} |L[\mathrm{embed}_{N,\eta}(b)] - L[b]| \leqslant 4M \left\{ \frac{1-\delta^\ell}{1-\delta}\, T\eta + a_N(\ell) \right\}. \tag{65}$$

Next, apply Theorem 5.1 (i) with $\mathrm{embed} = \mathrm{embed}_{N,\eta}$. The approximate Picard maps defining the Picard-type FNO are U_M-valued by (42); hence, the boundedness condition required in Theorem 5.1 is satisfied without adding an extra clipping layer. Thus, with probability at least $1-\zeta$,

$$L\left[\widehat{\Gamma}_{N,\ell,\eta}^{\mathrm{FNO}}\right] \leqslant \varepsilon_{\mathrm{imp}}^{\mathrm{FNO}}(N,\eta) + \frac{M^2\delta^{2\ell}}{(1-\delta)^2} + CM\widehat{\mathfrak{R}}^{\Omega_{\mathrm{eq}}} + CM^2\sqrt{\frac{\log(2/\zeta)}{n}}. \tag{66}$$

Here, the term $M^2\delta^{2\ell}/(1-\delta)^2$ is the oracle error coming from the Picard truncation estimate in Lemma 4.1. By combining (66) with (65) and Lemma F.8 (Lemma 7.3), we obtain (64). This proves the theorem. □

Corollary F.10 (Parameter choices for a prescribed risk)**.** *Assume the setting of Theorem F.9, and fix* $0 < \gamma < \min\{s_0 - d/2, 1\}$. *For any* $\varepsilon, \zeta \in (0,1)$*, the parameters can be chosen so that*

$$L\left[\widehat{\Gamma}_{N,\ell,\eta}^{\mathrm{FNO}}\right] \leqslant \varepsilon$$

with probability at least $1-\zeta$. *A sufficient choice has the scaling*

$$\eta \asymp \varepsilon, \quad H = O(\varepsilon^{-1}), \quad \ell = O\big(1 + \log \varepsilon^{-1}\big), \quad N = O(\varepsilon^{-1/\gamma}), \quad n = O\big(\varepsilon^{-2}(1+\log(1/\zeta))\big).$$

The implied constants may depend on $R, M, L, T, \delta, R_0, d, s_0, \gamma$*, but not on* ε, ζ*, or* q*. The number of retained Fourier modes satisfies*

$$|\Lambda_N| = (2N+1)^d = O(\varepsilon^{-d/\gamma}).$$

Proof. By Proposition F.6 and Theorem F.9, there exists a constant $C > 0$, depending only on the fixed parameters, such that

$$L\left[\widehat{\Gamma}_{N,\ell,\eta}^{\mathrm{FNO}}\right] \leqslant C\left(\eta + N^{-\gamma} + \delta^{2\ell} + \sqrt{\frac{1+\log(1/\zeta)}{n}}\right)$$

with probability at least $1-\zeta$. Choose the parameters so that each term on the right is at most $\varepsilon/4$. The network-size bound follows from $H(\rho_F) \leqslant C(1 + ML\eta^{-1})$. □

The bound does not require q to increase: $q = 1$ is sufficient for consistency as $n \to \infty$, with the approximation parameters chosen accordingly. This conclusion concerns the estimator obtained by empirical risk minimization over the abstract Picard class followed by the implementation map.

Remark F.11 (Fourier cutoff, depth, and the curse of parametric complexity). The bound (64) separates the three roles of the approximation parameters. The Picard depth ℓ appears in the geometric truncation term $M^2\delta^{2\ell}/(1-\delta)^2$. The scalar network accuracy η appears in the nonlinear implementation term $T\eta(1-\delta^\ell)/(1-\delta)$. The Fourier cutoff N appears only through $a_N(\ell)$. Proposition F.6 gives $a_N(\ell) \leqslant CN^{-\gamma}$ uniformly in ℓ for any $0 < \gamma < \min\{s_0 - d/2, 1\}$. Corollary F.10 gives sufficient parameter choices for a prescribed population risk. In contrast, the statistical term is controlled by the one-dimensional entropy of $\mathcal{F}_{M,L}$ and does not depend on the Fourier cutoff. This is the sense in which the Picard-type FNO implementation avoids the curse of parametric complexity at the statistical level, although the numerical cost of evaluating the Fourier representation still depends on the spatial dimension through the number of active modes.

F.6 Long-Time Prediction by Rolling Out the Picard-Type FNO

We finally record the long-time consequence of the local Picard-type FNO construction. This is a direct specialization of Theorem 6.1 to the torus setting. The terminal trace map

$$\psi : U_M \to C(\mathbb{T}^d), \quad \psi(u) := u(T)$$

is 1-Lipschitz with respect to the L^∞ norms:

$$\|\psi(u) - \psi(v)\|_{L^\infty(\mathbb{T}^d)} \leqslant \|u - v\|_{L^\infty(Q_T)}. \tag{67}$$

Thus, in the notation of Theorem E.1 (Theorem 6.1), we have $L_\psi = 1$ and $C_S = 1$. Let $\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta} : U_{0,R} \to U_M$ be a local Picard-type FNO model. For example, one may take the estimator constructed in Theorem 7.4, extended from U_0 to $U_{0,R}$ by the same Picard-type FNO formula. Define the exact local time-T transition by

$$\Phi_* := \psi \circ \mathcal{G}_*.$$

For $\kappa \in \mathbb{N}$, set

$$U_0^{[\kappa]} := \left\{ u_0 \in U_0 : \Phi_*^j(u_0) \in U_{0,R} \text{ for } j = 0, \dots, \kappa \right\}.$$

We assume $\Pi(U_0^{[\kappa]}) > 0$ and write $\Pi_\kappa := \Pi(\cdot \mid U_0^{[\kappa]})$.

For $u_0 \in U_0^{[\kappa]}$, define the exact rollout states by

$$v_0 := u_0, \quad v_{j+1} := \Phi_*(v_j), \quad j = 0, \dots, \kappa - 1.$$

Let $\mathcal{P}_R : C(\mathbb{T}^d) \to U_{0,R}$ be the pointwise clipping map

$$(\mathcal{P}_R f)(x) := \max\{-R, \min\{f(x), R\}\}.$$

The approximate rollout transition induced by the Picard-type FNO is

$$\widehat{\Phi}^{\mathrm{FNO}}_{N,\ell,\eta} := \mathcal{P}_R \circ \psi \circ \widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}. \tag{68}$$

The approximate states are defined by

$$\widehat{v}_0 := u_0, \quad \widehat{v}_{j+1} := \widehat{\Phi}^{\mathrm{FNO}}_{N,\ell,\eta}(\widehat{v}_j), \quad j = 0, \dots, \kappa - 1.$$

The clipping in (68) is used only to keep the next input state inside $U_{0,R}$. It is not an output clipping layer in the local Picard-type FNO; the local FNO itself is already U_M-valued by (42).

For $j = 0, \dots, \kappa - 1$, define the block-wise long-time risk by

$$L_j^{\mathrm{block}}\left[\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}\right] := \mathbb{E}_{u_0 \sim \Pi_\kappa,\, z \sim \mu}\left[\left|\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}(\widehat{v}_j)(z) - \mathcal{G}_*(v_j)(z)\right|^2\right].$$

Here, $z = (x, s) \in Q_T$ is the local space-time coordinate on the j-th block, corresponding to the global time $t = jT + s$.

Define the rollout-relevant state set by

$$\mathcal{V}_\kappa\left(\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}\right) := \left\{ v_j(u_0), \widehat{v}_j(u_0) : u_0 \in U_0^{[\kappa]},\ j = 0, \dots, \kappa \right\} \subset U_{0,R}.$$

The corresponding rollout-relevant local error is

$$\varepsilon^{[\kappa]}_{\mathrm{loc},N,\ell,\eta} := \sup_{v \in \mathcal{V}_\kappa(\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta})} \left\| \widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}(v) - \mathcal{G}_*(v) \right\|_{L^\infty(Q_T)}.$$

Theorem F.12 (Long-time rollout of the Picard-type FNO)**.** *Assume the setting of Appendix F.1 and let $\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta} : U_{0,R} \to U_M$ be a Picard-type FNO local model. Then, for any $\kappa \in \mathbb{N}$ and every $j = 0, \dots, \kappa - 1$,*

$$L_j^{\mathrm{block}}\left[\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}\right] \leqslant \left(\sum_{r=0}^{j} \left(\frac{1}{1-\delta} \right)^r \right)^2 \left(\varepsilon^{[\kappa]}_{\mathrm{loc},N,\ell,\eta} \right)^2. \tag{69}$$

In particular, the rollout step introduces no additional Rademacher complexity term. The block index j appears only through the deterministic stability factor in (69)*.*

Proof. This is Theorem 6.1 applied to the present torus setting. Indeed, by (34), the local solution operator satisfies

$$\|\mathcal{G}_*(u_0) - \mathcal{G}_*(v_0)\|_{L^\infty(Q_T)} \leqslant \frac{1}{1-\delta}\|u_0 - v_0\|_{L^\infty(\mathbb{T}^d)} \tag{70}$$

for all $u_0, v_0 \in U_{0,R}$. This is the estimate (4) with $C_S = 1$. Moreover, the terminal trace map satisfies $L_\psi = 1$ by (67). Hence, the stability factor in Theorem 6.1 becomes

$$L_\psi \frac{C_S}{1-\delta} = \frac{1}{1-\delta}.$$

Substituting this value into Theorem 6.1 gives (69). □

Remark F.13 (Oracle local-error specialization). If one considers the target-specific oracle implementation $\Gamma^{\mathrm{FNO}}_{N,\ell,\rho_{F_*}}$ and if the rollout-relevant states are contained in a fixed compact set $K_\kappa \Subset C(\mathbb{T}^d) \cap U_{0,R}$, then the compactness argument in the proof of Theorem 7.2, with U_0 replaced by K_κ, yields a Fourier error $a_{N,K_\kappa}(\ell) \to 0$ such that

$$\sup_{v\in K_\kappa} \left\|\Gamma^{\mathrm{FNO}}_{N,\ell,\rho_{F_*}}(v) - \mathcal{G}_*(v)\right\|_{L^\infty(Q_T)} \leqslant \frac{M\delta^\ell}{1-\delta} + \frac{1-\delta^\ell}{1-\delta}\,T\eta + a_{N,K_\kappa}(\ell).$$

Consequently, for this oracle Picard-type FNO,

$$L_j^{\mathrm{block}}\left[\Gamma^{\mathrm{FNO}}_{N,\ell,\rho_{F_*}}\right] \leqslant \left(\sum_{r=0}^{j}\left(\frac{1}{1-\delta}\right)^r\right)^2 \left(\frac{M\delta^\ell}{1-\delta} + \frac{1-\delta^\ell}{1-\delta}\,T\eta + a_{N,K_\kappa}(\ell)\right)^2.$$

This illustrates how the local Picard truncation error, scalar-network approximation error, and Fourier implementation error propagate over long time horizons. The Fourier rank still appears only through the local implementation error, and the rollout itself introduces no new statistical complexity term.

Remark F.14 (The case of defocusing nonlinearities). We record two useful refinements of the long-time rollout discussion for defocusing nonlinearities.

(a) Define the defocusing (or absorbing) subclass by

$$\mathcal{F}^{\mathrm{def}}_{M,L} := \{F \in \mathcal{F}_{M,L} : aF(a) \leqslant 0 \text{ for all } a \in [-M, M]\}.$$

Equivalently, $F(a) \leqslant 0$ for $a > 0$ and $F(a) \geqslant 0$ for $a < 0$. If $F_* \in \mathcal{F}^{\mathrm{def}}_{M,L}$, then the comparison principle gives

$$|u(t)| \leqslant e^{t\Delta_{\mathbb{T}^d}}|u_0|, \quad \|u(t)\|_{L^\infty(\mathbb{T}^d)} \leqslant \|u_0\|_{L^\infty(\mathbb{T}^d)}$$

for the exact solution of $\partial_t u - \Delta_{\mathbb{T}^d} u = F_*(u)$. Hence, the local mild solution extends globally in time. In particular, if $u_0 \in U_{0,R}$, then

$$\Phi_*^j(u_0) \in U_{0,R} \quad \text{for any } j \geqslant 0.$$

Therefore, for any finite rollout horizon $\kappa \in \mathbb{N}$, the admissible exact-rollout set satisfies

$$U_0^{[\kappa]} = U_0.$$

Thus, Theorem F.12 applies for arbitrarily long finite time horizons without imposing an additional admissibility assumption on the exact trajectory. This should not be confused with a uniform-in-time error bound: the deterministic factor in (69) may still grow with the block index j, and the rollout-relevant local error $\varepsilon^{[\kappa]}_{\mathrm{loc},N,\ell,\eta}$ may depend on κ.

(b) A stronger conclusion is available under strict dissipativity. For $\lambda > 0$, define

$$\mathcal{F}^{\mathrm{sd}}_{M,L,\lambda} := \left\{F \in \mathcal{F}_{M,L} : (F(a) - F(b))(a-b) \leqslant -\lambda|a-b|^2 \text{ for all } a, b \in [-M, M]\right\}.$$

If F is C^1, this condition is equivalent to $F'(a) \leqslant -\lambda$ on $[-M, M]$. In particular, $\mathcal{F}^{\mathrm{sd}}_{M,L,\lambda} \subset \mathcal{F}^{\mathrm{def}}_{M,L}$. Typical examples are

$$F(u) = -\alpha u - \beta|u|^{p-1}u, \quad \alpha \geqslant \lambda,\ \beta \geqslant 0,\ p \geqslant 1$$

(when $\alpha = 0$ and $p > 1$, the nonlinearity is defocusing but not strictly dissipative). Assume now that $F_* \in \mathcal{F}^{\mathrm{sd}}_{M,L,\lambda}$. Then the exact local solution operator is nonexpansive on each time block, and the exact terminal transition is a contraction:

$$\|\mathcal{G}_*(u_0) - \mathcal{G}_*(v_0)\|_{L^\infty(Q_T)} \leqslant \|u_0 - v_0\|_{L^\infty(\mathbb{T}^d)}, \tag{71}$$

and

$$\|\Phi_*(u_0) - \Phi_*(v_0)\|_{L^\infty(\mathbb{T}^d)} \leqslant q\|u_0 - v_0\|_{L^\infty(\mathbb{T}^d)}, \quad q := e^{-\lambda T} < 1. \tag{72}$$

Indeed, for two exact solutions u and v, the difference $w = u - v$ satisfies a linear equation with a zeroth-order coefficient bounded above by $-\lambda$, and the maximum principle yields the above estimates.

Consequently, the proof of Theorem F.12 can be sharpened as follows. Let

$$e_j := \|\widehat{v}_j - v_j\|_{L^\infty(\mathbb{T}^d)}.$$

Using (72) instead of the crude Lipschitz bound (70), we obtain

$$e_{j+1} \leqslant \varepsilon^{[\kappa]}_{\mathrm{loc},N,\ell,\eta} + qe_j, \quad e_0 = 0.$$

Hence

$$e_j \leqslant \frac{1-q^j}{1-q}\varepsilon^{[\kappa]}_{\mathrm{loc},N,\ell,\eta}.$$

Combining this with the block-wise nonexpansiveness (71), we get

$$L_j^{\mathrm{block}}\left[\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}\right] \leqslant \left(1 + \frac{1-q^j}{1-q}\right)^2 \left(\varepsilon^{[\kappa]}_{\mathrm{loc},N,\ell,\eta}\right)^2.$$

In particular,

$$L_j^{\mathrm{block}}\left[\widehat{\Gamma}^{\mathrm{FNO}}_{N,\ell,\eta}\right] \leqslant \left(1 + \frac{1}{1-e^{-\lambda T}}\right)^2 \left(\varepsilon^{[\kappa]}_{\mathrm{loc},N,\ell,\eta}\right)^2 \quad \text{for all } j = 0, \dots, \kappa - 1.$$

Thus, strict dissipativity replaces the growing stability factor in (69) by a factor uniformly bounded in the block index.

G Additional Remarks

Potential extensions. Although this paper mainly treats nonlinear parabolic PDEs, the essence of the framework lies not in parabolicity itself, but in the Picard-type fixed-point structure and in the ability to estimate the statistical complexity of the corresponding solution class. Picard iteration based on Banach's fixed point theorem is a standard method for constructing solutions to nonlinear evolution equations [Banach, 1922, Henry, 1981, Pazy, 1983, Zeidler, 1986, Cazenave and Haraux, 1998], and it also appears in the analysis of the Navier–Stokes equations, nonlinear dispersive equations, wave equations, and related problems [Giga and Miyakawa, 1985, Cannone, 2004, Cazenave, 2003, Tao, 2006]. Therefore, the framework developed in this paper may be extendable to a broader class of nonlinear evolution equations that admit a Picard-type fixed-point formulation.

Furthermore, even for problems where Picard iteration is not directly applicable, many iterative solvers are used in PDEs and inverse problems, such as Newton methods, monotone iteration methods, and iterative regularization methods [Deuflhard, 2011, Pao, 1992, Kaltenbacher and Huynh, 2022, Nguyen, 2024]. From this viewpoint, the idea of the present paper is not necessarily limited to Picard iteration, but may be extended to more general solver-aligned iterative frameworks. This direction is also consistent with recent work on neural general operator networks based on Banach fixed point iterations [Feischl et al., 2025], and such extensions are an important direction for future research.

Equation identification. The present paper uses the equation class Ω_{eq} as structural prior information, but it does not aim to identify the target equation specification ω_*. Although each $\omega \in \Omega_{\mathrm{eq}}$ determines a solution operator $\mathcal{G}_\omega$, recovering ω_* from data or from $\mathcal{G}_*$ is a separate inverse problem. This inverse direction raises questions of identifiability, stability, and observation design, and is beyond the scope of this paper. Extending the present framework toward equation identification is an interesting direction for future work.

Weight-tied architectures. The Picard-type constructions in this paper can also be viewed as weight-tied architectures. In Definition F.1, the same approximate Picard transition $\mathcal{T}^N_{\rho,u_0}$ is repeatedly applied, and the same scalar network ρ is shared across all Picard steps. The Fourier–heat operators S_N and $\mathcal{K}_N$ are fixed by the PDE structure. Thus increasing the Picard depth ℓ does not introduce independent layer-wise parameters, but corresponds to iterating the same PDE-aligned transition.

This is closely related to the weight-tied neural operators discussed in Furuya et al. [2025, Remark 3] and to fixed-point or equilibrium-type models. The present paper adds a statistical interpretation: because the Rademacher term is controlled by the abstract Picard transition class, increasing the number of Picard steps does not cause an unbounded growth of the entropy-based statistical term. This suggests that weight tying is not only memory-efficient, but also naturally aligned with the generalization theory of solver-aligned operator learning. A systematic empirical study of such weight-tied Picard-type neural operators is left for future work.

Beyond purely deterministic rollout. The long-time prediction result in this paper uses the learned local model without retraining it after the first training on short-time interval. This is consistent with the standard extension argument for nonlinear PDEs: once a local solution operator is available, the solution is extended by using the terminal value of one time block as the initial data for the next block. In this setting, the rollout step is deterministic and introduces no additional Rademacher complexity term.

From a practical viewpoint, however, this is only one possible strategy for long-time prediction. It is natural to consider correction mechanisms that reduce the accumulation of rollout errors. For example, one may introduce a learned correction term at each time block, retrain or fine-tune the local model along the rollout trajectory, or combine the learned model with data assimilation when partial observations are available. Such approaches may improve long-time stability in practice, especially for weakly stable or unstable dynamics.

These extensions require a different theoretical treatment. Once correction, retraining, or data assimilation is introduced during the rollout, the error is no longer obtained only by deterministic propagation of a fixed local model. One must also control the statistical error, optimization error, or assimilation error generated at each correction step. Developing a generalization theory for such corrected or data-assimilated rollouts is an important direction for future work.

Further PDE-specific refinements. In Section 7 and Appendix F.1, we treated semilinear heat equations on the torus as one concrete example of the general theory. This example was chosen because the heat semigroup is explicitly diagonalized by the Fourier basis and the Fejér cutoff gives an L^∞-stable implementation. However, the same strategy is not limited to the torus setting. Similar results should be obtainable for other domains, boundary conditions, linear operators, or basis expansions, provided that the corresponding semigroup estimates, compactness properties, and implementation errors can be controlled.

Moreover, different PDE classes carry different structural properties, and these properties may lead to sharper or more specialized bounds. For example, maximum principles, comparison principles, dissipativity, smoothing effects, spectral decay, or conservation laws can be incorporated into the rollout or implementation analysis when they are available. As a simple illustration, in Remark F.14 we discussed defocusing and strictly dissipative nonlinearities. In the defocusing case, the comparison principle gives global admissibility of the exact rollout, while in the strictly dissipative case the stability factor in the long-time error propagation can be replaced by a uniformly bounded one. These examples suggest that PDE-specific structure can be used not only to construct Picard-type implementations, but also to improve the resulting long-time prediction bounds. A systematic study of such PDE-specific refinements is left for future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the theoretical scope of the paper: Picard-type operator learning for nonlinear parabolic PDEs, the associated error decomposition, finite-observation bounds, rollout analysis, and the semilinear heat equation example.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations section discusses the dependence on contraction, boundedness, rollout stability, and architecture-specific implementation error estimates.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in the main text before the relevant results, and the paper points the reader to the appendix for the corresponding proof details.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: This paper does not include experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is a theoretical study of operator learning and does not involve human subjects, private data, deceptive systems, or high-risk deployment.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader Impacts section discusses potential benefits for reliable PDE surrogate modeling and the risk of overreliance on learned surrogates outside the assumptions of the theory.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with high misuse risk, such as pretrained generative models or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external code, datasets, or pretrained models as research assets; related scientific works are cited in the bibliography.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, code packages, or model checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Technical details are double-checked by LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.